Numerical computation of band gaps in photonic crystal fibres

Norton, Richard

Award date:
2008

Awarding institution:
University of Bath

Numerical Computation of Band Gaps in Photonic Crystal Fibres

submitted by

Richard Norton

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

September 2008

SUMMARY

Photonic crystal fibres are capable of special light guiding properties that ordinary optical fibres do not possess, and efforts have been made to numerically model these properties. The plane wave expansion method is one of the numerical methods that has been used. Unfortunately, the function that describes the material in the fibre $n(\mathbf{x})$ is discontinuous, and convergence of the plane wave expansion method is adversely affected by this. For this reason, the plane wave expansion method may not be every applied mathematician's first choice method but we will show that it is comparable in implementation and convergence to the standard finite element method. In particular, an optimal preconditioner for the system matrix A can easily be obtained and matrix-vector products with A can be computed in $\mathcal{O}(N \log N)$ operations (where N is the size of A) using the Fast Fourier Transform. Although we are always interested in the efficiency of the method, the main contribution of this thesis is the development of convergence analysis for the plane wave expansion method applied to 4 different 2nd-order elliptic eigenvalue problems in $\mathbb{R}$ and $\mathbb{R}^2$ with discontinuous coefficients.

To obtain the convergence analysis three issues must be confronted: regularity of the eigenfunctions; approximation error with respect to plane waves; and stability of the plane wave expansion method. We successfully tackle the regularity and approximation error issues but proving stability relies on showing that the plane wave expansion method is equivalent to a spectral Galerkin method, and not all of our problems allow this. However, stability is observed in all of our numerical computations.

It has been proposed in [40], [53], [63] and [64] that replacing the discontinuous coefficients in the problem with smooth coefficients will improve the plane wave expansion method, despite the additional error. Our convergence analysis for the method in [63] and [64] shows that the overall rate of convergence is no faster than before.

To define A we need the Fourier coefficients of $n(\mathbf{x})$, and sometimes these must be approximated, thus adding an additional error. We analyse the errors for a method where $n(\mathbf{x})$ is sampled on a uniform grid and the Fourier coefficients are computed with the Fast Fourier Transform. We then devise a strategy for setting the grid-spacing that will recover the convergence rate of the plane wave expansion method with exact Fourier coefficients.

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge the support and constructive criticism of my PhD supervisor, Robert Scheichl, whose enthusiasm and energy has helped propel this project towards completion from the beginning right through to the very end.

I would also like to thank the other staff and students at the University of Bath who have helped me along the way. In no particular order they are: David Bird, Greg Pearce, Ilia Kamotski, Vladimir Kamotski, Ivan Graham, Valery Smyshlyaev, John Toland, Geoffrey Burton, Adrian Hill, Alastair Spence, Jan Van Lent, Melina Freitag, Stefano Giani, Fynn Scheben and Nathan Broomhead.

The financial support of the Department of Mathematical Sciences at the University of Bath and an ORS award from Universities UK made this thesis possible.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 The Subject of the Thesis

Photonic Crystal Fibres (PCFs) are the next generation of optical fibre and physicists are actively trying to discover and exploit their unique optical properties. Because making PCFs is difficult and expensive, the task of mathematically modeling the behaviour of light in PCFs is important. In this thesis we consider the problem of computing band gaps and guided modes in PCFs using the plane wave expansion method. This is the same method that is used by physicists in the Centre for Photonics and Photonic Materials at the Physics Department of the University of Bath, [62], [63], [64] and [66].

The propagation of light is governed by Maxwell's equations, therefore, to model PCFs we need to solve Maxwell's equations. A commonly used approach when modeling PCFs is to make assumptions on the form of solutions based on the symmetries in the structure of the PCF and derive a formulation that is simpler than the full system of Maxwell's equations. It is important to realise that within PCF literature there are many different formulations of Maxwell's equations that authors use to model PCFs depending on the properties of the PCF they would like to model and the type of numerical method they would like to use. In this thesis we focus on four particular formulations of Maxwell's equations that are suited to the plane wave expansion method, although we also review other formulations that are used in the literature. The four formulations of Maxwell's equations that we consider are all linear second-order elliptic eigenvalue equations posed on $\mathbb{R}^d$, $d = 1, 2$, with coefficient functions that may be periodic and either piecewise constant or derivatives of piecewise constant functions. The four formulations that we consider are:

1. the *Full 2D Problem*, which is a 2D vector-valued eigenproblem;

2. the *Scalar 2D Problem*, which can be thought of as a simplified version of the

Full 2D Problem, although it is physically relevant in its own right under certain conditions; and

3. the *1D TE and TM Mode Problems*.

Both the Scalar 2D Problem and the 1D TE Mode Problem resemble Schrödinger's equation with a periodic, piecewise constant potential, whereas the Full 2D Problem and the 1D TM Mode Problem have an additional 1st-order term where the coefficient is a derivative of a periodic, piecewise constant function.

The correct mathematical framework to consider each of the eigenvalue equations is to define an equivalent operator on an appropriate Hilbert space. Our goal is to compute the spectra of these operators. Before we apply the plane wave expansion method, we exploit the periodicity of the coefficients in our operator by applying the Floquet Transform. This leads to a family of new differential operators over a bounded domain (the period cell) with periodic boundary conditions, which is crucial in order to apply the plane wave expansion method. A result from Floquet theory links the spectrum of our original operator to the spectra of our family of new operators. Moreover, the spectrum of each of our new operators is discrete.

Thus, our problem reduces to calculating the spectrum of a differential operator on a bounded domain using the plane wave expansion method. For example, consider the operator

$$L = \nabla^2 + V(x)$$

operating on $L^2_p(\Omega)$ where $\Omega = (-\frac{1}{2}, \frac{1}{2})^d$, $V(\mathbf{x}) \in L^2(\Omega)$ and $L^2_p(\Omega)$ is a function space that consists of functions in $L^2(\Omega)$ with periodic boundary conditions. Under additional regularity assumptions, finding λ in the spectrum of L is equivalent to finding an eigenpair (λ, u) such that

$$Lu = \lambda u \qquad \text{on } \Omega \tag{1.1}$$

where $u : \Omega \to \mathbb{C}$ satisfies periodic boundary conditions. To apply the plane wave expansion method to this eigenvalue equation we expand the eigenfunction $u(\mathbf{x})$ as a linear combination of plane waves,

$$u(\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{Z}^d} c_{\mathbf{k}} \, \mathrm{e}^{i2\pi \mathbf{k} \cdot \mathbf{x}} \tag{1.2}$$

for constants $c_{\mathbf{k}}$. For $d = 1$ we recognise (1.2) as the Fourier Series of $u(\mathbf{x})$. We also expand the coefficient function $V(\mathbf{x})$ in terms of plane waves (denoting the Fourier coefficients of $V(\mathbf{x})$ by $[V]_{\mathbf{k}}$). We then substitute (1.2) and our expansion of $V(\mathbf{x})$ into (1.1) to obtain,

$$-\sum_{\mathbf{k} \in \mathbb{Z}^d} |2\pi \mathbf{k}|^2 c_{\mathbf{k}} \, \mathrm{e}^{i2\pi \mathbf{k} \cdot \mathbf{x}} + \sum_{\mathbf{k}, \mathbf{k}' \in \mathbb{Z}^d} [V]_{\mathbf{k}'} c_{\mathbf{k}} \, \mathrm{e}^{i2\pi (\mathbf{k} + \mathbf{k}') \cdot \mathbf{x}} = \lambda \sum_{\mathbf{k} \in \mathbb{Z}^d} c_{\mathbf{k}} \, \mathrm{e}^{i2\pi \mathbf{k} \cdot \mathbf{x}} . \tag{1.3}$$

To get an approximation for the unknown eigenfunction $u(\mathbf{x})$ and its associated eigenvalue λ, we truncate the sum over $\mathbf{k} \in \mathbb{Z}^d$ to $|\mathbf{k}| \leq G$ (where G is a chosen integer), and then try to find the unknown eigenvalue λ and the unknown coefficients $c_{\mathbf{k}}$ with $|\mathbf{k}| \leq G$. We do this by matching the coefficients of the $\mathrm{e}^{i2\pi\mathbf{k}\cdot\mathbf{x}}$ terms for each $\mathbf{k}$ with $|\mathbf{k}| \leq G$. In this way we obtain a system of N (where N is the number of vectors $\mathbf{k} \in \mathbb{Z}^d$ with $|\mathbf{k}| \leq G$) linear equations for $N+1$ unknowns, which is equivalent to a matrix eigenproblem,

$$\mathrm{A}\,\mathbf{u} = \lambda\mathbf{u} \tag{1.4}$$

where $\mathbf{u} = (c_{\mathbf{k}})_{|\mathbf{k}|\leq G}$ is an N-vector of unknown coefficients and λ is the unknown eigenvalue in (1.3).

This matrix eigenproblem is then solved using whichever numerical technique is most appropriate for our needs. For all of our problems we will use a Krylov subspace iteration method as our eigensolver (since we do not need to compute all of the eigenvalues of A) and at each iteration of the eigensolver we will solve linear systems of the form $\mathrm{A}\,\mathbf{x} = \mathbf{b}$ using an iterative method (PCG or GMRES depending on whether or not A is symmetric positive definite). Inside our iterative linear solver we need to compute matrix-vector multiplications with A. The great advantage of the plane wave expansion method for all of our problems is that the operation of matrix-vector multiplication with A can be computed in $\mathcal{O}(N \log N)$ operations using the Fast Fourier Transform.

In the physics literature the plane wave expansion method for solving (1.1) is usually presented as we have just presented it; see for example [39] and [64]. In this thesis, to help with the error analysis, we will attempt to write the plane wave expansion method as a Galerkin method. Instead of solving a problem like (1.1) we will initially phrase the problem as a variational eigenvalue problem: Find an eigenpair (λ, u) such that $\lambda \in \mathbb{C}$, $0 \neq u \in \mathcal{H}$ and

$$a(u, v) = \lambda b(u, v) \qquad \forall v \in \mathcal{H} \tag{1.5}$$

where $\mathcal{H}$ is a suitable space of periodic functions and $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are bilinear forms. We apply the Galerkin method to (1.5) by introducing the finite dimensional subspace $\mathcal{S}_G \subset \mathcal{H}$, that is the span of a finite number of plane waves,

$$\mathcal{S}_G = \mathrm{span}\{\mathrm{e}^{i2\pi\mathbf{k}\cdot\mathbf{x}} : \mathbf{k} \in \mathbb{Z}^d, |\mathbf{k}| \leq G\}$$

and approximate (1.5) with the following discrete variational eigenproblem: Find an eigenpair (λ_G, u_G) such that $\lambda_G \in \mathbb{C}$, $u_G \in \mathcal{S}_G$ and

$$a(u_G, v_G) = \lambda_G b(u_G, v_G) \qquad \forall v_G \in \mathcal{S}_G. \tag{1.6}$$

For some of the problems we consider it is easy to show that the matrix eigenproblem

that we obtain from the plane wave expansion method is equivalent to a problem with the form of (1.6).

To estimate the error in the eigenvalues and eigenfunctions of (1.6) we use the theory in [6]. In this theory the errors in the approximate eigenfunctions and eigenvalues are analysed by studying the convergence of the corresponding solution operators. For example, we define the solution operator $\mathrm{T} : \mathcal{H} \to \mathcal{H}$ that corresponds to (1.5) by

$$a(\mathrm{T}\, u, v) = b(u, v) \qquad \forall v \in \mathcal{H}, u \in \mathcal{H}$$

and we define the solution operator $\mathrm{T}_G : \mathcal{H} \to \mathcal{S}_G$ that corresponds to (1.6) by

$$a(\mathrm{T}_G\, u, v_G) = b(u, v_G) \qquad \forall v_G \in \mathcal{S}_G, u \in \mathcal{H}.$$

Using the theory in [6] we bound the errors in the approximate eigenfunctions and eigenvalues in terms of

$$\| \mathrm{T}\, u - \mathrm{T}_G\, u \| \tag{1.7}$$

where $u(\mathbf{x})$ is a normalised eigenfunction of (1.5) and $\| \cdot \|$ is the energy norm induced by $a(\cdot, \cdot)$. We then use standard error analysis results for the Galerkin method to bound (1.7) in terms of the approximation error of $u(\mathbf{x})$ in $\mathcal{S}_G$, i.e. we bound (1.7) in terms of

$$\inf_{\chi \in \mathcal{S}_G} \|u - \chi\|.$$

Finally, to obtain the dependence of the approximation error on G (and thus on the number of degrees of freedom in our discrete problem) we need some further information about the regularity of the eigenfunctions of (1.5). Since our problems have coefficients that are not infinitely differentiable, the eigenfunctions of (1.5) have limited regularity. Therefore, the approximation error of the exact eigenfunctions in $\mathcal{S}_G$ does not decrease exponentially with G, and thus the plane wave expansion method does not converge exponentially with respect to G either.

In [40], [53], [63] and [64], the authors suggest that replacing the discontinuous (or derivatives of discontinuous) coefficient functions of our problem with smooth approximations of the coefficient functions will improve the plane wave expansion method. In this thesis we replicate the method in [63] and [64] (we call it the *smoothing method*) and we examine the two error contributions, from smoothing and from the plane wave expansion method. Our aim will be to extract the explicit dependence of the errors on the smoothing parameter as well as on the number of plane waves. To do this we will need to use Strang's 1st Lemma in a non-standard way as well as developing new regularity results.

When the structure of the PCF is relatively simple we can write down explicit formulae for the entries of the matrix A in (1.4), but for more complicated PCF structures

it is necessary to approximate the entries of A. Instead of using quadrature to do this, we will use an extremely efficient method called the *sampling method*. The method samples the values of a coefficient function on a uniform grid and then computes an approximation to the Fourier coefficients of the coefficient function by applying the Fast Fourier Transform. Again, we will use Strang's 1st Lemma to examine the error that the sampling method introduces.

As we have already discussed, to solve (1.4) we will use an iterative eigensolver as well as an iterative linear solver, and the Fast Fourier Transform to efficiently compute matrix-vector products with A. Another factor that influences the efficiency of the method is the number of iterations required by our linear solver. To reduce this we precondition the linear system so that instead of solving $\mathrm{A}\,\mathbf{x} = \mathbf{b}$, we solve

$$(\mathrm{P}^{-1}\,\mathrm{A})\mathbf{x} = \mathrm{P}^{-1}\,\mathbf{b} \tag{1.8}$$

where P is our preconditioner. It is another particular advantage of the plane wave expansion method that choosing P as the diagonal of A is a very effective preconditioner. If P is the diagonal of $\mathrm{A} + \mathrm{K}\,\mathrm{I}$ where K is a constant then (provided K is sufficiently large) we can bound the condition number of $\mathrm{P}^{-1}(\mathrm{A} + \mathrm{K}\,\mathrm{I})$ independently of G and N and numerical computations show that the number of iterations required to solve (1.8) remains constant as G and N increase. To ensure that the number of iterations required by our eigensolver is also independent of G and N we will actually choose P to be a block-diagonal part of A. This will ensure that we do not need to choose a large shift K.

1.2 The Aims of the Thesis

Associated with the plane wave expansion method, as with any numerical method, are errors. This thesis, being a thesis in numerical analysis, is dedicated to understanding and estimating the errors that arise from using the plane wave expansion method for band gap and guided mode computations in PCFs. We would like to show, using both theory and example, how the errors depend on the parameters of both the problem and the numerical method. A secondary issue that we also consider is an efficient implementation of the method.

The motivation for studying the problem of computing band gaps and guided modes in PCFs comes from a PhD thesis from the Physics Department at the University of Bath, [62], [63], [64], [66], where the plane wave expansion method and variations of the plane wave method have been used to compute band gaps in PCFs. To the best of our knowledge, only [8] and [79] have examined the errors of the plane wave expansion method for PCF problems. Purely based on numerical examples, they demonstrate that the plane wave expansion method is plagued by slow error convergence for these

types of problems. There does not appear to be any work in the literature that presents any mathematical error analysis of the plane wave expansion method applied to PCF problems. This thesis attempts to fill this gap.

In [63] and [64] the authors advocate the use of the smoothing method to improve the plane wave expansion method and to restore the exponential (or at least superalgebraic) convergence that one might expect for problems with infinitely differentiable coefficients. This claim seems to be dubious because smoothing introduces an additional error. We would like to carefully analyse the error contributions from both the smoothing and the plane wave expansion method so that we can answer the question: Is smoothing worth it?

The sampling method is also used in [64] in conjunction with the smoothing method for problems when the structure of the PCFs is complicated. This introduces an additional error. We would like to devise an optimal strategy for choosing the sampling grid-spacing so that the convergence rate of the plane wave expansion method without sampling can be recovered.

1.3 The Achievements of the Thesis

The main achievements of this thesis can be summarised as follows.

1. A complete error analysis of the standard plane wave expansion method applied to the Scalar 2D Problem and the 1D TE Mode Problem. This includes:

   (a) proving regularity results for the eigenfunctions of these problems;

   (b) showing that the eigenfunction error is optimal in the sense that we can bound it in terms of the approximation error;

   (c) bounding the approximation error in terms of the number of degrees of freedom in our finite dimensional subspace;

   (d) showing the eigenvalue error converges at twice the rate of the eigenfunction error; and

   (e) verifying with numerical examples that our error bounds are sharp (up to algebraic order).

   Ultimately, we show that the convergence of the plane wave expansion method depends on the regularity of the eigenfunctions. Since the problems that we consider have discontinuous coefficients, the regularity is limited, and therefore the convergence is also limited. This is why we do not see superalgebraic convergence of the plane wave expansion method.

2. A complete error analysis of the smoothing method applied to the Scalar 2D Problem and the 1D TE Mode Problem. This includes:

(a) bounding the error introduced by smoothing the coefficients of the original problem in terms of a smoothing parameter, by applying Strang's 1st Lemma in a non-standard way;

(b) proving regularity results for the problem with smoothed coefficients, determining explicitly the dependence on the smoothing parameter;

(c) using the regularity results to show that the plane wave expansion method converges superalgebraically for the smooth problem;

(d) showing that our eigenfunction error bounds are sharp (up to algebraic order) with numerical examples; and

(e) balancing the error contributions from smoothing and from the plane wave expansion method to obtain a strategy for choosing the amount of smoothing that minimises the error.

We show that the proposition in [64] that smoothing will improve the plane wave expansion method is false for the Scalar 2D Problem and the 1D TE Mode Problem when we have explicit formulae for the Fourier coefficients of the coefficient functions. Although we obtain superalgebraic convergence to the smooth solution, this is balanced by the additional error that is introduced by smoothing. The total error converges at the same rate as when no smoothing is applied.

3. A complete error analysis of the sampling method applied to the Scalar 2D Problem and the 1D TE Mode Problem. This includes:

   (a) bounding the error between a discontinuous function and its approximation via the sampling method;

   (b) applying Strang's 1st Lemma to obtain the additional error contribution from sampling;

   (c) demonstrating with numerical examples that our theoretical error bounds are correct (but not necessarily sharp) with numerical examples; and

   (d) balancing the error contributions from sampling with the plane wave expansion method errors to obtain a strategy for choosing the grid-spacing of sampling grid.

   We show that sampling, although it is a very efficient method because it allows us to calculate all of the Fourier coefficients with only one application of the Fast Fourier Transform, has a significant error contribution. This additional error can be mitigated by choosing a very fine sampling grid according to our strategy. Sometimes, however, the additional cost of our strategy is unfeasible and the error of the plane wave expansion method (without smoothing) can not be recovered.

4. An error analysis of the smoothing and sampling methods applied simultaneously. This includes choosing a strategy for setting the smoothing and sampling parameters that will minimise the errors. We put this strategy into practice with numerical examples.

5. An original result that proves that for the Scalar 2D Problem and the 1D TE Mode Problem, preconditioning with the diagonal of A (from (1.4)) is optimal in the sense that the condition number of A multiplied by the preconditioner is bounded independently of the size of A. This result is verified numerically by observing that the number of iterations required by our linear solver does not depend on the size of the linear system.

6. An error analysis of the standard plane wave expansion method and the spectral Galerkin method applied to the 1D TM Mode Problem. This includes:

   (a) proving regularity results for the exact eigenfunctions of the 1D TM Mode Problem;

   (b) using the regularity to bound the approximation error of exact eigenfunctions in terms of the degrees of freedom in our finite dimensional subspace;

   (c) complete error analysis for the spectral Galerkin method;

   (d) rewriting the plane wave expansion method as a non-conforming Petrov-Galerkin method (unfortunately, this does not lead to a stability result); and

   (e) observing through numerical examples that the plane wave expansion method is stable.

   Although we do not manage to prove a complete error analysis of the plane wave expansion method applied to the 1D TM Mode Problem, we successfully prove many of the necessary results. In particular, we prove a regularity result from which we derive an approximation error estimate. Numerical observations are consistent with the approximation error and we observe that the plane wave expansion method is stable for our numerical examples (even though we can not prove it). We also present the spectral Galerkin method for the 1D TM Mode Problem. Unlike for the 1D TE Mode Problem, the spectral Galerkin method is not the plane wave expansion method in this case. In contrast to the plane wave expansion method we can prove a complete error analysis for the spectral Galerkin method but we do not have an efficient implementation.

7. Numerically observed convergence rates for smoothing and sampling within the plane wave expansion method applied to the 1D TM Mode Problem.

8. Analysis of the existence of eigenpairs and of the regularity of eigenfunctions for the Full 2D Problem. This includes:

   (a) proving the existence of eigenpairs for the problem posed in 3D;

   (b) proving a regularity result for the eigenfunctions of the 3D problem;

   (c) proving the equivalence between the 2D and 3D problems;

   (d) using the regularity result in 3D to prove a regularity result for the 2D problem; and

   (e) showing that our regularity results are consistent with error calculations from numerical examples.

   The Full 2D Problem can be thought of (in a certain sense) as an extension of the 1D TM Mode Problem to 2D. Although we manage to prove many of the results for the Full 2D Problem that we proved for the 1D TM Mode Problem, the proof techniques are not the same and we are required to consider the full 3D system of Maxwell's equations in order to make any progress.

9. Numerically observed convergence rates for smoothing and sampling within the plane wave expansion method applied to the Full 2D Problem.

1.4 The Structure of the Thesis

The remainder of this thesis is divided into five chapters.

In Chapter 2 we give the physical background for PCFs and we discuss, in detail, the different mathematical models that can be derived from Maxwell's equations to model PCFs. We review the extent to which each of the models has been studied in the literature, with particular emphasis on the mathematical analysis for each model and on the various numerical methods that have been applied to the different models.

In Chapter 3 we review the many and varied mathematical tools that we will require for the error analysis and for the implementation of the plane wave expansion method. Some of these results are original and interesting in their own right. We begin with some preliminary definitions of function spaces and mollifiers. Throughout this thesis we will be working with periodic functions and this is the topic of the next section in Chapter 3. In particular, we define periodic Sobolev spaces and we present several results about their properties. The next section is on piecewise continuous functions. Of particular importance in this section is the regularity result for a special class of piecewise continuous functions. We then present some definitions and results from spectral theory, including the definition of the Floquet Transform. Following the spectral theory we present some results from functional analysis. Within this section we include a key result from [6] for the error analysis of the Galerkin method applied to

a variational eigenvalue problem. We also present Strang's 1st Lemma as well as a few regularity results for elliptic boundary value problems. Finally, we present a section on numerical linear algebra including the tools for solving (1.4).

In Chapter 4 we present the bulk of our error analysis contribution for the plane wave expansion method. In this chapter we consider both the Scalar 2D Problem and the 1D TE Mode Problem. We begin by correctly (in the spectral theory sense) presenting the problem as that of calculating the spectrum of an operator on a Hilbert space. We apply the Floquet Transform, and then the plane wave expansion method. We include the implementation details and we prove a result about a possible preconditioner before presenting our error analysis. Finally, we consider the smoothing and sampling methods for these problems.

In Chapter 5 we consider two methods applied to the 1D TM Mode Problem: the plane wave expansion method and the spectral Galerkin method (which are not equivalent for the 1D TM Mode Problem). We begin by writing the problem as an operator on a Hilbert space and applying the Floquet transform, from which we obtain a variational eigenproblem to solve. We then present a section on the implementation of the plane wave expansion method. Following the implementation details we consider the error analysis for the plane wave expansion method and we begin by proving a result about the regularity of the eigenfunctions for the exact problem. Our first attempt at the error analysis is to use the same techniques that we used in Chapter 4, by applying the spectral Galerkin method to our variational eigenproblem. This approach is successful in obtaining a complete error analysis, but the spectral Galerkin method is not the plane wave expansion method for the 1D TM Mode Problem and it does not have the same implementation efficiencies that the plane wave expansion method has. Instead, we show that the plane wave expansion method is equivalent to a non-conforming Petrov-Galerkin method. Unfortunately, we are unsuccessful in completely analysing the error for this problem. Using our regularity result for the eigenfunctions of the exact problem we derive an approximation error result and this gives us an upper limit for the rate at which the plane wave expansion method can converge for the eigenfunctions. We then observe that the plane wave expansion method actually achieves this optimum rate of convergence for some numerical examples. We also provide numerical examples of smoothing and sampling within the plane wave expansion method.

In Chapter 6 we consider the Full 2D Problem. Without being able to appropriately phrase the problem as an operator on a Hilbert space we are limited to following the technique in [64] to present the plane wave expansion method. We do, however, manage to prove a regularity result by considering an equivalent problem in 3D from which we can determine the regularity of eigenfunctions of the 2D problem. Using this regularity result we can derive an approximation error estimate for plane waves approximating

an eigenfunction of the 2D problem. Since our approximation error result measures the best possible approximation of an eigenfunction using plane waves, it provides us with an upper limit for the rate at which the plane wave expansion method can converge for eigenfunctions. Numerical examples show that this upper limit is actually achieved by the plane wave expansion method for these examples, and thus, it is the regularity of the exact problem that is limiting the convergence rate of the plane wave expansion method.

CHAPTER 2
PHYSICS

In this chapter we discuss Photonic Crystal Fibres (PCFs) from a physical perspective and we introduce the mathematical model that is used to study PCFs. We begin by giving a physical description of what PCFs are and what physical properties we would like them to have. We support this discussion with references for applications of PCFs. We then introduce the mathematical model for the interaction of light with PCFs, based on Maxwell's equations. We make assumptions (based on the symmetries in PCFs) on the form of the solution and manipulate Maxwell's equations to arrive at the formal equations that we wish to solve. Following the formulation of equations that model PCFs we present a review of results on the mathematical analysis of these equations. This is followed by a review of the many numerical methods that have been applied to solving the various formulations of Maxwell's equations for PCFs. A key reference for this chapter is [64].

2.1 Description of PCFs

Traditional optical fibres that are in use in the communications industry guide light by a phenomenon known as *total internal reflection*, [76]. This occurs when light travels in a material of high refractive index and is confined to the material by a series of reflections at the interface with a low refractive index material. If the direction of the incident light makes a sufficiently acute angle with the interface then all of the light is reflected back into the high refractive index material. PCFs guide light by a different physical phenomenon and it is this different physical phenomenon that we want to model mathematically.

Before we describe PCFs we must first discuss photonic crystals. Photonic crystals were first proposed by Yablonovitch [90] and John [41]. Just as electrons can be manipulated by periodicity of an atomic lattice in a semiconductor crystal (to get energy ranges over which no allowed electronic states exist), Yablonovitch and John

proposed the existence of crystals for which propagation of certain frequency ranges of light through the crystal would be forbidden. Semiconductors have *electronic band gaps* where certain electronic states do not exist, whereas photonic crystals have *photonic band gaps* where there is a range of light frequencies for which propagation through the crystal is forbidden. We make the distinction between 1D, 2D and 3D photonic crystals depending on how many directions the crystal varies in. Figure 2-1 has a diagram of a 1D photonic crystal where the crystal only varies in the vertical direction.

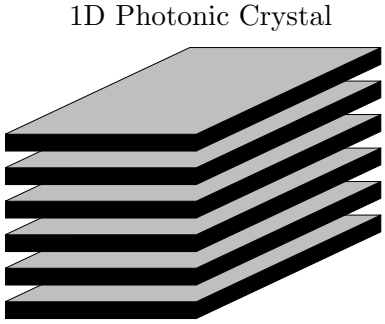

Figure 2-1: Diagram of a 1D photonic crystal.

Now we describe PCFs. A PCF is a long thin cylinder of 2D photonic crystal (that varies in the transverse/cross-sectional directions only) with a defect running down the centre of the cylinder, see Figure 2-2. We refer to the central defect as the *core* of the fibre and the surrounding 2D photonic crystal as the *cladding*. We align axes so that the z-axis runs along the core of the PCF and the transverse coordinates are x and y. Theoretically, the structure of PCF is invariant along the length of the fibre, however, true invariance is impossible to manufacture. For our modelling purposes we will assume that the PCF is constant with respect to the z-direction. Typically, PCFs are made from silica with air holes running along the length of the fibre. A regular periodic array of air holes in the cross-section of the fibre forms the 2D photonic crystal in the cladding of the fibre whereas the core of the fibre is a defect in the crystal structure, usually formed by either the absence of one or more air holes in the centre of the fibre or an especially large air hole in the centre. PCFs with a large air hole in the core of the fibre are called *hollow core* PCFs and we only consider PCFs of this type in this thesis. The shape, size and pattern of air holes in the cladding, as well as the shape and size of the core, of PCFs varies between fibres and contribute towards their photonic properties. The material used to make PCFs also influences the photonic properties.

The aim is to manufacture a PCF so that there exists a mode of light (i.e. light of a specific frequency) that is guided along the centre of the fibre. We call this a *guided*

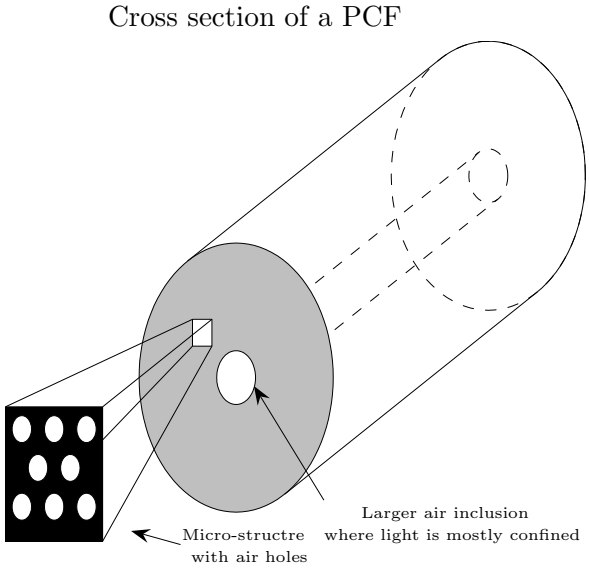

Figure 2-2: Diagram of a PCF.

mode. For this to be achieved the cladding of the fibre must act as a barrier and forbid the propagation of this particular mode through it.

For this reason let us first model the pure 2D photonic crystal that is found in the cladding. We must find a *band gap* for the 2D photonic crystal - a range of light frequencies where propagation through the photonic crystal is forbidden.

Once we have found a band gap in the 2D photonic crystal we have a clue as to where we might try to find a guided mode in the corresponding PCF. Since the 2D photonic crystal has a band gap, we expect the cladding of our PCF to act as a barrier to light with frequencies from this band gap. Therefore, a guided mode should have a frequency from this band gap. However, we can not be sure that a guided mode will be permitted in the core of the PCF from studying the 2D photonic crystal. We must also model the entire PCF to find possible guided modes. This idea is supported by analysis which we discuss later in this chapter.

We call the PCFs we have considered so far 2D PCFs since the photonic crystal in the cladding of the fibre is a 2D photonic crystal. We can also consider 1D PCFs. There are two ways we can construct these. The first is to consider a 1D photonic crystal made from slabs with a planar defect, i.e. a defect that is only confined in the direction in which the photonic crystal varies. The second way is to construct a fibre with a central defect running along the core of a fibre where the cladding is a 1D photonic crystal that varies only in the radial direction. The second construction is referred to as a Bragg fibre [91].

For an introduction to PCFs and their applications please refer to two popular review articles, [48] and [70], or the book [39].

2.2 Formulation of Equations

In this section we formulate the equations that model the interaction of light and PCFs. Light is a form of electro-magnetic radiation and is governed by Maxwell's equations. To formulate equations for modeling PCFs we make assumptions on how the electric and magnetic fields depend on t (time) and z (the spatial coordinate running along the length of the fibre). These assumptions are based on the symmetries in PCFs and are the same assumptions that are made in [76] page 590 and 591, for example. Taking advantage of these assumptions, to reformulate Maxwell's equations, yields a 2D vectorial eigenproblem. This will form the core problem to be solved and analysed in this thesis.

However, we also derive other systems of equations that have been used in the literature to model PCFs. We do this to draw attention to the difference between our model and the models used by others. In particular, we highlight that an additional assumption is needed to decouple the full 2D vectorial problem that we solve into two scalar problems, as it is often done in the mathematical literature. In this case, the two scalar problems are polarised such that either the electric or magnetic field are entirely in the directions transverse to the z-axis. Our full model is not restricted by this additional assumption.

Although we do not solve the decoupled scalar problems mentioned above, we will use other simplified models where appropriate to develop a deeper theoretical understanding of PCFs and the numerical methods we use to solve PCF problems.

We also consider 1D PCFs. In this case we make an additional assumption and the equations naturally decouple into scalar equations.

We begin with source-free Maxwell's equations for a non-magnetic material. The system of equations is

$$\nabla \cdot (n^2 \mathbf{E}) = 0 \tag{2.1}$$

$$\nabla \cdot \mathbf{H} = 0 \tag{2.2}$$

$$\nabla \times \mathbf{H} = \epsilon_0 n^2 \frac{\partial \mathbf{E}}{\partial t} \tag{2.3}$$

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \tag{2.4}$$

where $\mathbf{E}$ is the electric field vector, $\mathbf{H}$ is the magnetic field vector, ϵ_0 is the permittivity of free space ($8.854 \times 10^{-12} F m^{-1}$), μ_0 is the permeability of a vacuum ($4\pi \times 10^{-7} N A^{-1}$) and n is the refractive index of the material. n completely describes the physical

properties of the PCFs; for 2D PCFs $n = n(x, y)$ and for 1D PCFs $n = n(x)$ (where we assume that 2D PCFs are aligned with the z-axis so that the crystal structure in the cladding varies in the x and y directions and 1D PCFs are aligned so that the crystal structure varies in the x direction). Alternatively, Maxwell's equations can be formulated in terms of the dielectric function or electric permittivity ϵ instead of the refractive index n. There is no fundamental difference in these formulations because $\epsilon = \epsilon_0 n^2$ and ϵ_0 is a constant.

For 2D PCFs we will refer to the directions that are perpendicular to the z-axis as the transverse directions.

2.2.1 Time Harmonic Maxwell's Equations

The first assumption that we make is that we assume (as in almost all photonics literature, eg. [76] and [39]) that the electric and magnetic fields can be written as $\mathbf{E}(\mathbf{x}, t) = \mathrm{e}^{-i\omega t}\widetilde{\mathbf{E}}(\mathbf{x})$ and $\mathbf{H}(\mathbf{x}, t) = \mathrm{e}^{-i\omega t}\widetilde{\mathbf{H}}(\mathbf{x})$ where ω is a specified frequency. More general solutions to Maxwell's equations can then be recovered by taking linear combinations of solutions of this type. With this representation of $\mathbf{E}$ and $\mathbf{H}$ we get

$$\frac{\partial \mathbf{E}}{\partial t} = -i\omega\mathbf{E}, \qquad \frac{\partial \mathbf{H}}{\partial t} = -i\omega\mathbf{H}$$

and (2.1)-(2.4) become source-free, non-magnetic, time harmonic Maxwell's equations

$$\nabla \cdot (n^2\widetilde{\mathbf{E}}) = 0 \tag{2.5}$$

$$\nabla \cdot \widetilde{\mathbf{H}} = 0 \tag{2.6}$$

$$\nabla \times \widetilde{\mathbf{H}} = -i\epsilon_0 n^2 \omega \widetilde{\mathbf{E}} \tag{2.7}$$

$$\nabla \times \widetilde{\mathbf{E}} = i\mu_0 \omega \widetilde{\mathbf{H}}. \tag{2.8}$$

We proceed by substituting (2.7) into (2.8) to get

$$\nabla \times \left(\frac{1}{n^2}\nabla \times \widetilde{\mathbf{H}}\right) = k_0^2 \widetilde{\mathbf{H}}$$

where $k_0 := \sqrt{\epsilon_0\mu_0}\omega$ is called the wave number. Alternatively, we could substitute (2.8) into (2.7) and obtain

$$\nabla \times \nabla \times \widetilde{\mathbf{E}} = k_0^2 n^2 \widetilde{\mathbf{E}}$$

To solve Maxwell's equations for a 3D photonic crystal problem we would need to solve either

$$\nabla \times \left(\frac{1}{n^2}\nabla \times \widetilde{\mathbf{H}}\right) = k_0^2 \widetilde{\mathbf{H}} \tag{2.9}$$

$$\nabla \cdot \widetilde{\mathbf{H}} = 0 \tag{2.10}$$

or

$$\nabla \times \nabla \times \widetilde{\mathbf{E}} = k_0^2 n^2 \widetilde{\mathbf{E}} \tag{2.11}$$

$$\nabla \cdot (n^2 \widetilde{\mathbf{E}}) = 0. \tag{2.12}$$

In both cases k_0^2 is an eigenvalue for the system of equations. Sometimes (2.9) and (2.11) are written with ω as the eigenvalue.

2.2.2 Invariance in z-direction

The second assumption that we make (as in [76]) is that we can represent the electric and magnetic field by

$$\widetilde{\mathbf{E}}(\mathbf{x}) = \mathbf{e}(x,y)\,\mathrm{e}^{i\beta z} = (\mathbf{e}_t(x,y) + e_z(x,y)\hat{\mathbf{z}})\,\mathrm{e}^{i\beta z} \tag{2.13}$$

$$\widetilde{\mathbf{H}}(\mathbf{x}) = \mathbf{h}(x,y)\,\mathrm{e}^{i\beta z} = (\mathbf{h}_t(x,y) + h_z(x,y)\hat{\mathbf{z}})\,\mathrm{e}^{i\beta z} \tag{2.14}$$

where $\mathbf{h}_t$ and $\mathbf{e}_t$ are vector fields that point in the tranverse directions and β is the z-component of the wave vector (the term wave vector comes from the representation for a wave $A\exp(i\mathbf{k}\cdot\mathbf{x})$ where $\mathbf{k}$ is called the wave vector). Again, more general solutions to the Maxwell's equations can be obtained by taking linear combinations of solutions of this type.

Substituting this representation into (2.9) and using (2.10) together with the identity $\nabla(\frac{1}{n^2}) = -\frac{1}{n^2}\nabla(\log n^2)$ we get (after some vector calculus) the following two equations

$$(\nabla_t^2 + k_0^2 n^2)\mathbf{h}_t - (\nabla_t \times \mathbf{h}_t) \times (\nabla_t \log n^2) = \beta^2 \mathbf{h}_t \tag{2.15}$$

$$(\nabla_t^2 + k_0^2 n^2) h_z \hat{\mathbf{z}} - (i\beta\hat{\mathbf{z}} \times \mathbf{h}_t + \nabla_t \times h_z\hat{\mathbf{z}}) \times (\nabla_t \log n^2) = \beta^2 h_z \hat{\mathbf{z}} \tag{2.16}$$

where $\nabla_t := (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, 0)$. If we fix ω (so that k_0^2 is fixed) then (2.15) is a 2D complex-valued eigenproblem for an eigenfunction $\mathbf{h}_t = (h_x, h_y, 0)$ and an eigenvalue β^2. Moreover, given a solution to (2.15) the other components of the magnetic and electric field are given by

$$h_z = \frac{i}{\beta}\nabla_t \cdot \mathbf{h}_t \qquad \text{from (2.6)} \tag{2.17}$$

$$e_z = i\sqrt{\frac{\mu_0}{\epsilon_0}}\frac{1}{k_0^2 n^2}\hat{\mathbf{z}} \cdot \nabla \times \mathbf{h}_t \qquad \text{from (2.7)} \tag{2.18}$$

$$\mathbf{e}_t = -\sqrt{\frac{\mu_0}{\epsilon_0}}\frac{1}{k_0^2 n^2}\hat{\mathbf{z}} \times (\beta\mathbf{h}_t + i\nabla_t h_z) \qquad \text{from (2.7).}$$

In this thesis we are interested in solving (2.15) and we call this the *Full 2D Problem*.

Since n^2 is a discontinuous function $\nabla_t \log n^2$ is not defined in the classical sense,

so we must consider (2.15) formally and rephrase the problem in terms of an operator on an appropriate Hilbert space that corresponds to (2.15). To find band gaps and guided modes of PCFs we will investigate the spectrum of this operator.

Note that in the formulation above we have implicitly fixed the frequency ω (equivalent to fixing k_0) and the intention is to solve for β. The band gaps we seek will be band gaps of β and not ω. Alternative formulations fix β and search for k_0 (ω) in a 2D problem, or solve a 3D problem for eigenvalues k_0.

As well as solving (2.15) we will also consider solving a scalar 2D problem in this thesis. We obtain the scalar 2D problem by omitting the $(\nabla_t \times \mathbf{h}_t) \times (\nabla_t \log n^2)$ term from (2.15). The resulting equation can then be decoupled into an equation for h_x and an equation for h_y, both of which take the same form, namely

$$\nabla_t^2 h + k_0^2 n^2 h = \beta^2 h. \tag{2.19}$$

We call (2.19) the *Scalar 2D Problem*. In [7] the authors call this equation the *scalar wave equation* and they argue that it can be applied to PCFs that have low contrast n^2.

2.2.3 Splitting into TE and TM modes (2D) - special case $\beta = 0$

In this section we review a special case of the Full 2D problem. Although we will not use this approach in this thesis, it is important to mention it because it has received a lot of attention in the literature, especially in the mathematical literature. For example, see [5], [15], [45] and [26].

It is an example of a formulation where β is fixed and the intention is to solve for an eigenvalue k_0^2, but it only applies in the case $\beta = 0$. By assuming that $\beta = 0$ Maxwell's equations conveniently decouple into two scalar equations.

If we assume again (2.13) and (2.14) (with $\beta = 0$) and substitute $\widetilde{\mathbf{H}} = \mathbf{h}_t(x, y) + h_z(x, y)\hat{\mathbf{z}}$ into (2.9) and (2.10) and $\widetilde{\mathbf{E}} = \mathbf{e}_t(x, y) + e_z(x, y)\hat{\mathbf{z}}$ into (2.11) and (2.12), then some vector calculus reveals that the problem decouples into two scalar problems with solutions of the form $(\widetilde{\mathbf{H}}, \widetilde{\mathbf{E}}) = (0, 0, h_z, e_x, e_y, 0)$ and $(\widetilde{\mathbf{H}}, \widetilde{\mathbf{E}}) = (h_x, h_y, 0, 0, 0, e_z)$ where $\mathbf{e}_t = (e_x, e_y, 0)$ and $\mathbf{h}_t = (h_x, h_y, 0)$. We call these two polarisations the transverse electric (TE) mode and the transverse magnetic (TM) mode, respectively. The equation that governs the TE mode is the equation for the z-component in (2.9), i.e.

$$-\nabla_t \cdot \left(\frac{1}{n^2} \nabla_t h_z \right) = k_0^2 h_z. \tag{2DTE}$$

Given a solution for h_z and the fact that $\mathbf{h}_t = \mathbf{0}$ and $e_z = 0$, $\mathbf{e}_t$ is determined by (2.7).

The equation that governs the TM mode is the equation for the z-component in

(2.11), i.e.

$$-\nabla_t^2 e_z = k_0^2 n^2 e_z. \tag{2DTM}$$

$\mathbf{h}_t$ is determined using (2.8).

Note that choosing $\beta = 0$ is equivalent to considering waves that only propagate in the transverse directions (and not in the z-direction). Since we are interested in waves that will propagate along the core of the fibre the assumption that $\beta = 0$ is not appropriate for our model. For $\beta \neq 0$ Maxwell's equations do not decouple.

However, the assumption that $\beta = 0$ is appropriate when studying truly 2D photonic crystals. Our 2D PCFs are actually 3D structures and we have reduced Maxwell's equations to a 2D problem by exploiting symmetries. An example of a 2D photonic crystal is a plate that has had a 2D structure etched onto it. Propagation is only possible in the plane of the plate, and not through the plate. Therefore, the assumption that $\beta = 0$ is appropriate in this case.

2.2.4 1D problem

In this subsection we formulate equations for 1D PCFs. We make the assumption that $n = n(x)$ (i.e. the photonic crystal in the cladding of the 1D PCF only varies with respect to x) and that the magnetic (and electric) fields have $\mathrm{e}^{i\beta_y y}$ dependence (β_y is a constant). With these assumptions we reduce (2.15) to a decoupled system of scalar equations.

We first write

$$\widetilde{\mathbf{H}}(\mathbf{x}) = \mathbf{h}(x)\,\mathrm{e}^{i(\beta_y y + \beta z)}\,.$$

In fact, without loss of generality we can choose $\beta_y = 0$. This is possible by rotating the y and z coordinate axes and keeping the x axis unchanged to force $\beta_y = 0$. In this case equation (2.15) becomes the decoupled system

$$\frac{d^2 h_x}{dx^2} + k_0^2 n^2 h_x = \beta^2 h_x \tag{2.20}$$

$$\frac{d^2 h_y}{dx^2} + k_0^2 n^2 h_y - \frac{d(\log n^2)}{dx}\frac{dh_y}{dx} = \beta^2 h_y \tag{2.21}$$

where $\mathbf{h}_t = (h_x, h_y, 0)$. If we solve (2.20) for non-zero h_x and set $h_y = 0$ (which satisfies (2.21)) then $e_z = 0$ by (2.18). The solution has the form $(\widetilde{\mathbf{H}}, \widetilde{\mathbf{E}}) = (h_x, 0, h_z, e_x, e_y, 0)$ with the electric field normal to the z-axis. Therefore, we call (2.20) the transverse electric (TE) mode equation.

Conversely, if we solve (2.21) for non-zero h_y and set $h_x = 0$ (which satisfies (2.20)) then $h_z = 0$ by (2.17). The solution has the form $(\widetilde{\mathbf{H}}, \widetilde{\mathbf{E}}) = (0, h_y, 0, e_x, e_y, e_z)$ with the magnetic field normal to the z-axis. Therefore, we call (2.21) the tranverse magnetic (TM) mode equation.

Just as for the Full 2D Problem in (2.15), the term $\frac{d(\log n^2)}{dx}$ in (2.21) is not defined in the classical sense. We must consider (2.21) formally and consider the problem as an operator on an appropriate Hilbert space. In this case we have been successful at rewriting the equation and we can write (2.21) in divergence form. Using the identity $-\frac{d(\log n^2)}{dx} = n^2 \frac{d}{dx}(\frac{1}{n^2})$ we can rewrite (2.21) as

$$\frac{d}{dx}\left(\frac{1}{n^2}\frac{dh_y}{dx}\right) + k_0^2 h_y = \frac{\beta^2}{n^2} h_y. \tag{2.22}$$

This form of (2.21) will be useful for the numerics and the analysis later in this thesis.

2.2.5 Boundary Conditions/Defining n on all of $\mathbb{R}^2$

So far we have not yet discussed the domains and boundary conditions for our eigenproblems. If we are trying to model a pure (infinite) phontonic crystal then n is periodic and it is defined on all of $\mathbb{R}$ or $\mathbb{R}^2$, and the problem is well defined without specifying boundary conditions. In reality however, a PCF is of course bounded and n is defined on a bounded domain in $\mathbb{R}^2$. In order to make the problem well defined we need to specify a domain (which may be a subset of the set in which n is defined) and boundary conditions. Alternatively, we can extend n outside of our chosen domain to all of $\mathbb{R}^2$ or $\mathbb{R}$ and consider our eigenproblems on unbounded domains. First, we discuss the supercell method before considering other methods.

The most popular method and the method that we use in this thesis is the *supercell method*. In the supercell method, n is extended periodically to all of $\mathbb{R}^2$ or $\mathbb{R}$. The original PCF in a bounded domain is called the *super cell* (see right pane of Figure 2-3). After applying the supercell method we have an eigenvalue problem with periodic coefficients posed on an unbounded domain. By using the Floquet-Bloch transform we exploit this periodicity and we transform the problem into a family of problems on bounded domains with periodic boundary conditions. The periodic boundary conditions are crucial for applying the plane wave expansion method. Examples of the supercell method for PCF problems can be found in [62], [64], [66] and [78]. For an example of the supercell method applied to a non-photonics problem, see [61].

A second technique for defining n on all of $\mathbb{R}^2$ (or $\mathbb{R}$) is to define it by extending the cladding of n to all of $\mathbb{R}^2$ (or $\mathbb{R}$). The overall structure is then an infinite 2D (or 1D) photonic crystal with a localised defect (see left pane of Figure 2-3). This technique for defining n on all of $\mathbb{R}^2$ (or $\mathbb{R}$) is commonly used in mathematical analysis literature because the classical Weyl theorem (at least for the 1D TE Mode Problem and the Scalar 2D Problem) states that the addition of a compact perturbation (localised defect) does not change the essential spectrum of the operator. Therefore, there is a clear connection between the spectrum of a "PCF" with this structure and the spectrum of a pure (infinite) photonic crystal. Unfortunately, this technique does not lead to

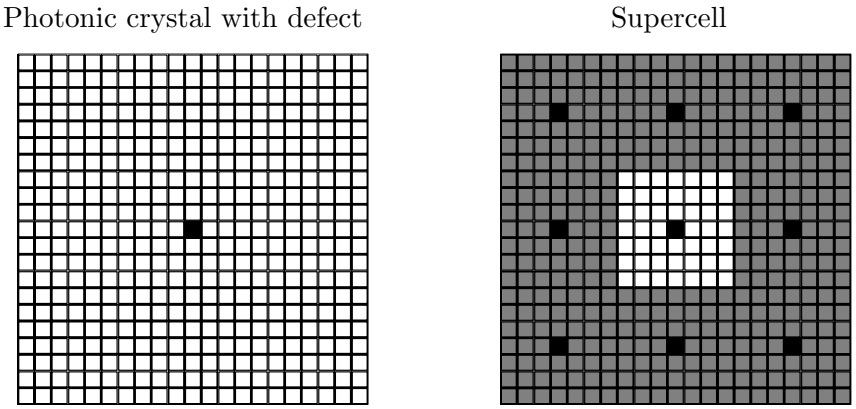

Figure 2-3: Diagram showing structure of n for two different choices of method for extending n to all of $\mathbb{R}^2$. The period cell of the supercell is highlighted.

a problem on a bounded domain (unlike the supercell method where we could use the Floquet-Bloch transform) and so it is not well suited for any numerical method. However, efforts have been made to design an exact absorbing boundary condition for this situation [29].

Another method for providing boundary conditions is given in [14]. In this paper the authors give boundary conditions for solving the Scalar 2D Problem on a bounded domain that are equivalent to extending n with $n(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbb{R}^2$ with $|\mathbf{x}| > R$ where R is the radius of the PCF.

All three of the techniques we have just described contain some form of modeling error because we do not know what boundary conditions represent reality. However, since we are searching for guided modes, and these should decay exponentially in the cladding, it is argued that the particular choice of boundary conditions (or how we extend n) is irrelevant provided there is a sufficient amount of cladding around the central defect. Moreover, the location of band gaps (in which we search for guided modes) can be calculated by considering a pure (infinite) photonic crystal.

In this thesis we need to impose periodicity on the coefficients of our problems (so that we can apply the plane wave expansion method) and we do this by applying the supercell method. We would like to have a theoretical justification that the supercell method does not introduce an excessive amount of error. Soussi's paper [78] links the supercell method to the infinite photonic crystal with a localised defect (second technique given above) for the special case of the decoupled 2D problems. He shows that the error in the essential spectrum between the photonic crystal with a localised defect and the supercell method decays quadratically with the inverse of the distance between neighbouring defects and that the error of isolated eigenvalues (guided modes)

decays exponentially with the distance between neighbouring defects, i.e. the more cladding between the defects in a supercell lattice, the less effect artificially introduced defects in the supercell lattice have.

The link between the supercell method and a (pure infinite) photonic crystal with a localised defect for the problems that we will study (1D TE and TM Mode Problems, Scalar 2D Problem and Full 2D Problem) has not yet been considered in the mathematical literature. However, we expect that similar results to those in [78] apply for all of our problems, and we observe this for a 1D TE Mode Problem example. In Figure 2-4 we have plotted the errors in the spectrum of a 1D TE Mode Problem between the supercell method and a photonic crystal with a localised defect and we observe that the error in the essential spectrum decays quadratically with the inverse of the number of cells in the cladding, while the errors in the discrete spectrum (isolated eigenvalues) decays exponentially. Our error calculations were made by solving Model Problem 2 in Chapter 4 with the plane wave expansion method for different numbers of cells in supercell cladding. To calculate the errors in the essential spectrum we have compared the spectrum of Model Problem 2 with the spectrum of a pure photonic crystal (i.e. the spectrum of Model Problem 1 in Chapter 4) because this remains unchanged when a localised defect is introduced. To calculate the errors in the discrete spectrum we notice that since all of the spectrum of a supercell operator is essential spectrum (because it has periodic coefficients), we find that there are narrow bands of essential spectrum of Model Problem 2 that approximate isolated eigenvalues. The discrete spectrum errors are the "widths" of these narrow bands.

For the rest of this thesis we will ignore the error introduced by the supercell method and concentrate on estimating the errors from the numerical methods that we apply to problems with periodic coefficients.

2.3 Overview of Analysis

In this section we give an overview of the results from the literature that apply to the problems that we have formulated in the previous section. The results that can be found in the literature are limited to the TE and TM mode problems in 1D and 2D, the *Scalar 2D problem*, and the 3D Maxwell problem in (2.9). There is no analysis in the literature of the *Full 2D problem* in (2.15) (although some progress has been made towards studying a scattering by diffraction problem that makes similar assumptions to ours).

For the formulations that have received attention in the literature, the analysis of each problem attempts to follow a common approach. First, the formal eigenvalue equation is considered as an operator on a Hilbert space. Then, for periodic n (modelling a perfect photonic crystal), the spectrum of the operator is found to be purely

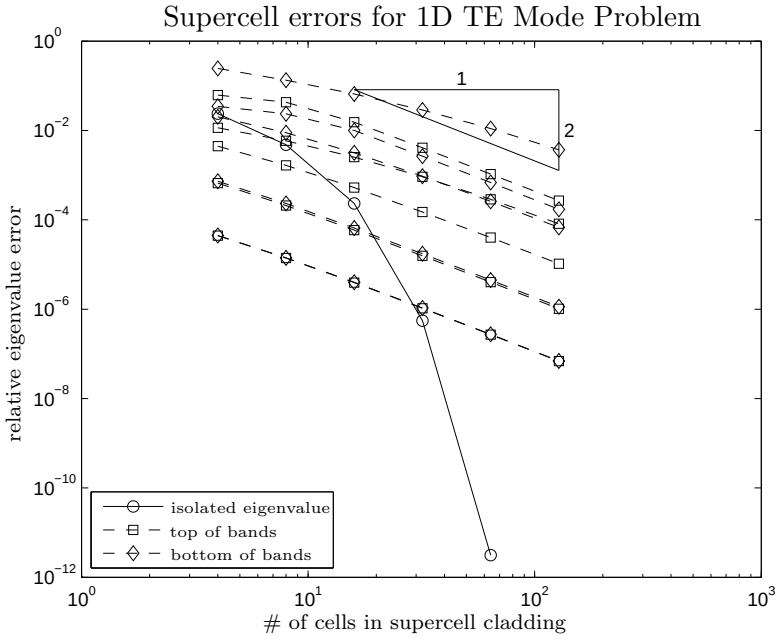

Figure 2-4: Plot of the relative error of the isolated eigenvalue and the bands for the 1D TE Mode Problem vs. the number of cells in the supercell cladding.

essential spectrum, and the existence of band gaps is proven. Next, a compact perturbation is added to n. With the addition of this compact perturbation it is proven that the essential spectrum is unchanged and for an eigenvalue with finite multiplicity in a band gap the corresponding eigenfunction must decay exponentially, i.e. we have a guided mode. However, some of these statements have not been proven for all of the above problems.

We note that the main tool for studying periodic operators is Floquet Theory (called Bloch theory in the physics literature). References for Floquet Theory include [17], [44], [45] and [69]. We discuss Floquet Theory in more detail in Chapter 3.

We also remark that it is often the case in the literature that authors have proved that the spectrum of an operator is *absolutely continuous* instead of working with the definition of essential spectrum. In Section 3.4.2 we give the definition of absolutely continuous spectrum that can be found in [42] where it is also stated that absolutely continuous spectrum is a subset of essential spectrum.

1D TE mode and Scalar 2D Problem

The Scalar 2D Problem in (2.19) is (mathematically speaking) the 2D extension of the 1D TE mode equation (2.20), and both equations are examples of Schrödinger's equation

$$-\nabla^2\psi + V(x)\psi = E\psi$$

identifying h with the wave function ψ, $k_0^2 n^2$ with the potential $V(x)$, and β^2 with the energy E.

According to Floquet Theory [45], the spectrum of periodic, ellipitic differential operators exhibit band structure and so the spectrum of the Schrödinger operator with periodic V will also exhibit band structure.

The following result for the 1D TE Mode Problem can be found in [69]: If V is periodic then the spectrum of the operator corresponding to (2.20) is absolutely continuous and if V is not constant then there must be gaps in the spectrum. This result is also known as Borg's Uniqueness Theorem. In 2D, a result in [69] states that if V has a Fourier Series where the coefficients are in l_2 (i.e $V \in L^2$) then the spectrum is absolutely continuous. The appearance of gaps for the 2D problem is not guaranteed for non-constant V but it is still a common occurance according to [45] and can be demonstrated numerically.

If we add a compact perturbation to V then it follows from the classical Weyl theorem (page 117 of [69]) that the essential spectrum remains unchanged. This means that any additional eigenvalues that appear must be of finite multiplicity. If such an eigenvalue appears in a band gap then it must decay exponentially in the cladding [45].

1D TM mode

The analysis of the 1D TM mode is covered in [25]. The operator corresponding to (2.22) is defined in terms of a quadratic form, for which the standard Floquet theory does not apply. In [25] the authors develop the corresponding Floquet theory that proves that the 1D TM mode has spectrum with band structure as well as proving sufficient conditions for the existence of band gaps. Perturbations of pure photonic crystal are not considered in [25].

Full 2D Problem

The Full 2D Problem (2.15) is not the 2D version of the 1D TM mode equation (2.21) and there are no papers in the literature that are dedicated to the spectral theory of this problem. We have had no success with rewriting the Full 2D problem in divergence form (as we did for the 1D TM mode problem in (2.22)) so that the coefficients are defined in the classical sense. Writing the Full 2D problem in an appropriate operator form remains an open problem. However, we use analytical results for the full 3D Maxwell operator to help describe the spectral properties of (2.15). We do this in Chapter 6.

Other analysis results that may be applicable to this problem, or may point the way forward in terms of how to approach the analysis of this problem, can be found in [19] and [20]. In these papers a conical diffraction problem is considered and the authors make similar assumptions to ours on the magnetic and electric fields before

reformulating Maxwell's equations in terms of the z-component of the magnetic and electric field in regions where n is constant together with interface conditions. They then prove existence, uniqueness and regularity results for their problem. However, they only assume that n is periodic in one of the coordinate directions and the results can not be directly applied to our problem.

2D TE and TM modes

Although we do not solve either the 2D TE or TM mode problem here, both of these problems have received a lot of attention in the literature. The band gap structure of the spectrum of these operators was established in [26]. A theorem for the absolute continuity of the TM mode for piecewise continuous, periodic n is given in [45]. However, absolute continuity of the spectrum of the TE mode has only been proven for smooth, periodic n, not piecewise continuous n [45]. [26] establishes the existence of band gaps for the TE and TM modes for square geometries where the appearance of gaps can be generated by increasing the size of the jump in n. Gaps in the TM mode spectrum for more general shaped geometries in n are studied in [27]. The corresponding article for the TE mode spectrum is referred to as being in preparation in [45] but it appears to have not been published. For a survey of these results, refer to [45].

We would like to emphasise again, however, that these problems assume that $\beta = 0$ and are therefore confined to waves that only propagate in the transverse directions.

Full 3D Maxwell System

Finally, let us consider the existing literature on the full 3D time-harmonic Maxwell operator corresponding to (2.9). The Hilbert space for this operator must be a subset of the vector fields that satisfy (2.10). The application of Floquet Theory to the Maxwell operator is not as straight forward as for elliptic operators with periodic coefficients, however, it is achieved by considering the Maxwell operator in an elliptic complex. See [46] and references therein for more details about this (in particular, see [24]). A consequence of the application of Floquet theory is that the spectrum has band structure. [59] proves that the spectrum of the Maxwell operator is absolutely continuous for smooth and periodic n^2, but not for discontinuous n^2. The existence of band-gaps has been verified with numerical experiments in, for example, [39]. [28] appears to be the only paper where the existence of a band gap has been proven, but this was for a hypothetical problem where $\mu \neq 1$ and there are high contrasts for n^2 and μ.

Localised defects are known not to change the essential spectrum of a photonic crystal (see Theorem 21 in [45] and references therein), but in 3D the defect in a PCF is a line defect. According to [45] there is no rigorous mathematical analysis of this problem although a relatively simple result that can be proven is that a mode with an

eigenvalue in a band gap must decay exponentially in the cladding. This can be proven by estimating the decay of the Green's function.

Analytic Solution to 1D Problems in a Photonic Crystal

The existence of an eigenvalue and eigenfunction for the 1D TE or 1D TM mode equations for a simple photonic crystal can be shown to be equivalent to finding a zero of a transcendental equation. We can use this to get an exact solution in 1D to compare our numerical results against.

The technique is to consider even and odd modes separately. The TE or TM mode equation is solved on each section of the period cell where n is constant and then the solutions are matched with appropriate interface conditions. An eigenfunction exists when the determinant of the coefficients is equal to zero. Expanding the determinant we obtain a transcendental equation that depends on the eigenvalue. By varying the eigenvalue we can find zeros of the transcendental equation that correspond to the existence of eigenpairs. This technique is explained in detail in the appendix of [64].

The 1D TE mode, as previously discussed, is just Schrödinger's equation and is called the Kronnig-Penney model when the potential is periodic. Solution techniques for this problem that are different from [64] are given in [55] and [23].

When the supercell method is applied then the period cell is more complicated than for a photonic crystal and the number of interface conditions to satisfy is much greater. In this case we resort to numerical methods to find a reference solution rather than deriving an expression for the determinant of the matrix of coefficients.

2.4 Overview of Numerical Methods

In this section we review the different numerical methods that have been applied to solving the PCF problem. Although we will focus on using the plane wave expansion method in this thesis there are many different methods that could be used to solve the PCF problem and they are often suited to particular formulations of Maxwell's equations.

Methods fit into one of two categories: frequency domain methods and time domain methods. Frequency domain methods are based on formulations of Maxwell's equations that are derived from the time-harmonic Maxwell equations while time domain methods are based on formulations of Maxwell's equations that include time dependence.

We begin with a review of the use of the plane wave expansion method for solving the PCF problem before briefly reviewing a number of other methods.

The review in [64] is more extensive and contains a review of various other methods used for solving the PCF problem.

Plane Wave Expansion Method

The plane wave expansion method is an example of a frequency domain method. For some problems it is equivalent to a Galerkin method. Sometimes it is referred to (as we do in this thesis) as a spectral Galerkin method. This is because the basis functions have global support. For PCF problems it is not a truly spectral method because the basis functions are not eigenfunctions of the operator. Another name for the method is the Fourier Galerkin method. It has been applied to all of the different formulations of Maxwell's equations with the only condition being that the coefficients are periodic. This condition is naturally satisfied for pure photonic crystals but is artificially imposed for PCFs using the supercell method.

Imposing periodicity in the coefficients introduces an error and prevents the plane wave expansion method from being able to model the effects of energy leaking through the cladding, i.e. *leaky modes*. However, since guided modes decay exponentially in the cladding, this error is small for guided modes. The non-localised modes that do not decay in the cladding and are not changed by the introduction of a localised defect can be dealt with by considering the simpler problem of solving the problem for the pure photonic crystal that corresponds to the cladding material. This issue was also discussed in Subsection 2.2.5.

The research group in the Physics Department at the University of Bath apply the plane wave expansion method where the frequency has been fixed and (2.15) is solved for the magnetic field and β, [62], [64] and [66]. In [53] the plane wave expansion method is applied to a 3D photonic crystal. Other examples of using the plane wave expansion method in PCFs include [38], [34], [15] and [40].

According to [79] the plane wave expansion method converges slowly for increasing numbers of plane waves and it is claimed that this is due to the discontinuous nature of the dielectric function. However, it is claimed in [75] and [8] that the slow convergence (for the 1D TM mode problem) is also influenced by how the plane wave expansion method is formulated for discontinuous data. The apparent slow convergence of the plane wave expansion method is essentially the phenomenon that we will attempt to understand in more detail in this thesis.

The advantages of the plane wave expansion method are that it is easy to formulate, and fast to compute, using the Fast Fourier Transform (FFT) and a preconditioner. The disadvantages are that it is apparently slow to converge when the data is discontinuous.

Two methods for improving the performance of the plane wave expansion method have been suggested in [64] and [63]. The first method they use is to replace the discontinuous coefficients with smooth coefficients that approximate the discontinuous coefficients. The smooth coefficients are obtained by convoluting the discontinuous coefficients with a normalized Gaussian function. Although this method may improve the convergence rate of the plane wave expansion method we must also consider the addi-

tional error that has been introduced. The analysis of smoothing is another important topic of this thesis.

The second method for improving the performance of the plane wave expansion method is to use curvilinear coordinates. When the structure of the discontinuous coefficients is complicated then for the plane wave expansion method we must approximate the Fourier coefficients of the discontinuous coefficients. A method that samples the discontinous coefficients on a uniform grid and then applies the Fast Fourier Transform is usually applied. However, to improve this approximation, the author of [64] has suggested sampling the discontinuous coefficients on a non-uniform mesh with nodes clustered near the discontinuities. Although we do not manage to analyse the error for this method in this thesis, we make the observation that this method lessens the effectiveness of the preconditioner that is used in [64].

Time Domain Methods

Time domain methods do not extract a $e^{i\omega t}$ dependence from the electric or magnetic fields as in the time harmonic Maxwell's equations. In these methods the solution to Maxwell's equations is propagated foward in time from some initial magnetic or electric field condition. The finite-difference time-domain (FDTD) method has been used in [68] for PCFs and is described in the books [82] and [49].

Once a solution has been computed with a time domain method the Fourier Transform of the solution then reveals peaks that correspond to the frequencies of the modes that propagate through the fibre. The disadvantage of FDTD methods is that the time dependent ODE system that is derived from spatial discretisation is stiff. This means that to preserve the stability of the ODE solver either the time step must decrease with the spatial grid spacing or an implicit time integrator must be used.

Beam Propagation Method

Beam propagation methods are another example of a frequency domain method, however, instead of computing guided modes they are used to compute propagation along a fibre. They begin by separating the z-dependence of the electric or magnetic field as $\Phi(x, y, z) = e^{i\omega z}\,\phi(x, y, z)$ where ω is a chosen frequency and $\phi(x, y, z)$ still depends on z, albeit in a slowly varying way. This is followed by discretisation in the transverse direction. The result is an ODE system that depends on z. The field (beam) is then propagated forward along the fibre in the z-direction using an ODE solver. There are a number of versions that use either finite difference, finite element or discrete Fourier transform discretisation schemes for the transverse direction discretisation. Examples of the beam propagation method applied to optical fibre problems are [71] and [22]. In [**?**], leaky modes are computed while in [22] a Fourier transform technique is described

for recovering information about guided and leaky modes that have been exited by the source beam.

Spectral Methods

The multipole method [88], [89] and the method in [23] are both examples of spectral methods. They construct basis functions that are orthogonal and are matched to the geometry of the PCF so that the discontinuities in n^2 will not affect the exponential convergence of the method. Both methods can only be applied to PCFs with particular geometries: eg. circular or square air holes.

In the multipole method time-harmonic Maxwell's equations are expressed in terms of the z-component of the magnetic and electric fields, ω is fixed and the equations are solved for β on a domain in the transverse directions. The method expands h_z and e_z in terms of basis functions that are the solution to the underlying equations in the different regions of the PCF where n is constant, which for a PCF with circular holes are cylindrical harmonics. If the PCF was constructed using some other geometrical shapes then different basis functions need to be used. The expansions of the solution in the different regions of the PCF are then matched at the interface between regions of different n as well as at the boundary of the domain.

The advantage of this method is that it is very efficient (because the discontinuities of n do not effect the convergence rate) and it is possible to model *leaky modes* (where some modes are only partially guided).

However, a disadvantage of this method is that it is limited by the range of PCF structures that it applies to. In practice it has only been applied to PCFs with circular holes. Another disadvantage of this method is that it is relatively difficult to implement.

Finite difference / finite element / boundary element / localised Gaussian-Hermite

All of these methods are standard methods that have been applied in the frequency domain by solving equations based on the time-harmonic Maxwell equations. They require setting a boundary condition on a bounded domain and they can be applied to PCFs of arbitrary geometry.

The finite difference method is applied to the *Full 2D problem* in [11]. The finite difference discretisation scheme leads to an eigenvalue problem where the matrix is sparse and banded. A method of of reordering the matrix elements is used to reduce the matrix bandwidth and then a subspace iteration method is used to find only a few of the eigenvalues of the matrix. The authors demonstrate that their method is significantly faster than the method used in [40].

The finite element (FE) method is applied to the 2D TE and TM mode problems in [15] and [5]. A uniform grid is used in [15] while an unstructured grid is used in [5]. The

uniform grid approach of [15] is easy to implement, and a preconditioner that utilises the Fast Fourier Transform (FFT) is used. The disadvantage of the method in [15] is that the rectangular uniform grid is necessary for their preconditioner, since it uses the FFT, and so elements cannot be concentrated in the regions where n is discontinuous (i.e. where the solution has less regularity). The method in [5] uses an unstructured mesh and a method called simultaneous coordinate over-relaxation is used to solve the matrix eigenproblem that arises from the FE discretisation. Both of these FE methods only solve the 2D TE and TM mode equations for photonic crystals. They do not solve the problems for PCFs.

Another PhD thesis at the University of Bath by Stefano Giani [31] also solves the 2D TE and TM mode problems using the FE method. Giani's work extends the FE method to the PCF problem and he uses *a posteriori* error estimation to refine the mesh in areas where the residual error is large.

For the boundary element method see [33] and [86].

Examples of localised Gaussian-Hermite methods are found in [56] and [58]. The method is similar to the plane wave expansion method and the finite element method except that the solution is expanded in terms of localised Gaussian-Hermite functions.

2.5 Summary of Problems

Let us summarise the eigenproblems that we will consider in this thesis. We write the problems in *dimensionless* form. Define Λ as the *lattice pitch*, i.e. Λ is the width of a period cell in the photonic crystal. Then we scale to get the following problems with $\lambda = \lambda_0\Lambda$, $\tilde{\beta} = \beta\Lambda$,

$$\gamma(\mathbf{x}) = \frac{4\pi^2}{\lambda_0^2} n^2(\mathbf{x}\Lambda)$$
$$\eta(\mathbf{x}) = \log n^2(\mathbf{x}\Lambda).$$

In this way we can rescale our eigenproblems so that the periodic coefficients have periodicity 1. In later chapters we will make further restrictive assumptions on the coefficients.

The four problems we consider in this thesis are then described by the following.

Problem 2.1 (**Full 2D Problem).** The primary problem we are interested in is the *Full 2D Problem* (2.15),

$$(\nabla_t^2 + \gamma(\mathbf{x}))\mathbf{h}_t - (\nabla_t \times \mathbf{h}_t) \times (\nabla_t \eta(\mathbf{x})) = \tilde{\beta}^2 \mathbf{h}_t$$

for 2D vector eigenfunctions $\mathbf{h}_t$ and eigenvalues $\tilde{\beta}^2$.

Problem 2.2 (**Scalar 2D Problem).** A secondary problem we are interested in is the *Scalar 2D Problem* (2.19),

$$\nabla_t^2 h + \gamma(\mathbf{x})h = \tilde{\beta}^2 h$$

for scalar eigenfunctions h and eigenvalues $\tilde{\beta}^2$.

In 1D, with the same scaling and definitions of γ and η we solve the following problems.

Problem 2.3 (**1D TE Mode Problem).** The *1D TE Mode Problem* is (2.20),

$$\frac{d^2h}{dx^2} + \gamma(x)h = \tilde{\beta}^2 h$$

which is an eigenproblem for scalar eigenfunction h and eigenvalue $\tilde{\beta}^2$.

Problem 2.4 (**1D TM Mode Problem).** The *1D TM Mode Problem* is (2.21)

$$\frac{d^2h}{dx^2} + \gamma(x)h - \frac{d\eta}{dx}\frac{dh}{dx} = \tilde{\beta}^2 h$$

which is an eigenproblem for scalar eigenfunction h and eigenvalue $\tilde{\beta}^2$.

CHAPTER 3
MATHEMATICAL TOOLS

In this chapter we develop the mathematical tools needed for the analysis of the plane wave expansion method applied to band gap computations in photonic crystal fibres. We split the chapter into six sections. In Section 3.1 we define a variety of function spaces, including test functions and distributions. We also introduce mollifiers and we present a lemma for estimating series in terms of integrals. In Section 3.2 we present some definitions and results for periodic functions and periodic distributions. In particular, we define finite dimensional periodic function spaces as well as various projections onto these function spaces. These will be important for presenting the plane wave expansion method as a Galerkin method. In Section 3.3 we develop results for describing the regularity of piecewise continuous functions. In Section 3.4 we present some results from spectral theory and Floquet theory. Section 3.5 has some results from functional analysis. It describes the abstract tools that are necessary for studying variational eigenvalue problems and it includes the main theorem that we use for the error analysis of the Galerkin method applied to a variational eigenvalue problem. We also present Strang's First Lemma in this section as well as some regularity results for elliptic boundary value problems. Finally, in Section 3.6, we present the tools from numerical linear algebra that we need for solving matrix eigenvalue problems.

3.1 Preliminaries

In this section we make some preliminary definitions. We begin by defining the function space $L^p_{loc}(\mathbb{R}^d)$ for $1 \leq p \leq \infty$. We then develop distributions in the standard way before defining the spaces $H^s(\mathbb{R}^d)$ and $H^s(\Omega)$ for $s \in \mathbb{R}$ in terms of the Fourier transform. Next, we define the standard mollifier and finally, we present a lemma for estimating series in terms of integrals.

Throughout this thesis $d \in \mathbb{N}$, although sometimes we restrict d so that $d \in \{1, 2\}$.

Bold letters, such as $\mathbf{x}$, will denote vectors in $\mathbb{R}^d$. A vector $\mathbf{x} \in \mathbb{R}^d$ will have entries $x_1, x_2, \ldots, x_d$ and we define $\mathbf{x}' := (x_1, x_2, \ldots, x_{d-1}) \in \mathbb{R}^{d-1}$. If $d = 3$ then we will sometimes use the notation $\mathbf{x}_t = (x_1, x_2, 0)$ (t for *transverse*) and $\mathbf{x}_z = (0, 0, x_3)$.

A vector $\alpha = (\alpha_1, \ldots, \alpha_d)$ with non-negative integer entries α_i is called a multi-index. The order of a multi-index is $|\alpha| := \alpha_1 + \cdots + \alpha_d$ and the factorial of α is defined as $\alpha! = \alpha_1!\alpha_2! \ldots \alpha_d!$.

We will use the following notation for partial derivative operators

$$D^\alpha := D_{x_1}^{\alpha_1} \ldots D_{x_d}^{\alpha_d} := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$$

and for $\mathbf{x} \in \mathbb{R}^d$ we denote

$$\mathbf{x}^\alpha := x_1^{\alpha_1} \ldots x_d^{\alpha_d}.$$

The support of a function $f : \mathbb{R}^d \to \mathbb{C}$ is defined as

$$\operatorname{supp} f := \overline{\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \neq 0\}}.$$

The open ball with centre $\mathbf{x} \in \mathbb{R}^d$ and radius $r > 0$ is denoted by

$$B(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^d : |\mathbf{x} - \mathbf{y}| < r\}.$$

Throughout this thesis we will be working with inequalities to estimate certain quantities. To avoid defining a large number of constants we will use the following notation: If $\frac{C}{D}$ is bounded above independent from our discretization parameters n, G, N, M, Δ then we write $C \lesssim D$. We will also write $C \simeq D$ when $C \lesssim D$ and $C \gtrsim D$.

We will use the Kronecker-delta symbol to denote the following function, for $i, j \in \mathbb{Z}$,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

For two functions $f, g : \mathbb{R} \to \mathbb{R}$ we write $f(x) = \mathcal{O}(g(x))$ (as $x \to \infty$) if there exist constants $C > 0$ and $x_0 > 0$ such that $|f(x)| \leq C|g(x)|$ for all $x > x_0$. Alternatively, we may write $f(x) = \mathcal{O}(g(x))$ as $x \to 0$ if there exist constants $C > 0$ and $x_0 > 0$ such that $|f(x)| \leq C|g(x)|$ for all $0 \leq x < x_0$. In these situations we say that f has order g.

Throughout this thesis we will use the term *superalgebraic* convergence (as $n \to \infty$) to mean that the error is $\mathcal{O}(n^{-s})$ for all $s \in \mathbb{R}$.

3.1.1 The Space $L^p_{loc}(\mathbb{R}^d)$

The function space $L^p_{loc}(\mathbb{R}^d)$ for $1 \leq p \leq \infty$ is defined as

$$L^p_{loc}(\mathbb{R}^d) := \{f|_K \in L^p(K) : \text{for any compact } K \subset \mathbb{R}^d\}$$

where $L^p(K)$ is defined in the usual way.

3.1.2 Test Functions and Distributions

In this subsection we define distributions in the usual way. Let $\Omega \subseteq \mathbb{R}^d$ be an open set.

Definition 3.1. Define the space of test functions on Ω as

$$\mathcal{D}(\Omega) = C^\infty_0(\Omega) = \{\phi \in C^\infty(\Omega) : \operatorname{supp}\phi \text{ is a compact subset of } \Omega\}.$$

Convergence in $\mathcal{D}(\Omega)$ is defined as follows: Let $\{\phi_n\}_{n=1}^\infty \subset \mathcal{D}(\Omega)$ be a sequence of test functions and let $\phi \in \mathcal{D}(\Omega)$. We say ϕ_n *converges to* ϕ *in* $\mathcal{D}(\Omega)$ and write $\phi_n \xrightarrow{\mathcal{D}} \phi$ as $n \to \infty$ if the following properties hold

1. there exists a compact set $K \subset \Omega$ such that $\operatorname{supp}\phi_n \subset K$ for all $n \in \mathbb{N}$.
2. $\max_{\mathbf{x}\in\Omega} |D^\alpha(\phi_n(\mathbf{x}) - \phi(\mathbf{x}))| \to 0$ as $n \to \infty$, for any multi-index α.

We now use this definition of $\mathcal{D}(\Omega)$ and convergence in $\mathcal{D}(\Omega)$ to define distributions.

Definition 3.2. A linear functional $u : \mathcal{D}(\Omega) \to \mathbb{R}$ is a distribution on Ω if

$$\phi_n \xrightarrow{\mathcal{D}} \phi \quad \Longrightarrow \quad \langle u, \phi_n\rangle \to \langle u, \phi\rangle$$

for any convergent sequence of test functions. The space of all distributions on Ω is denoted by $\mathcal{D}'(\Omega)$. A sequence $\{u_n\}_{n=1}^\infty \subset \mathcal{D}'(\Omega)$ converges to $u \in \mathcal{D}'(\Omega)$ if

$$\langle u_n, \phi\rangle \to \langle u, \phi\rangle \qquad n \to \infty,\ \forall \phi \in \mathcal{D}(\Omega).$$

Every $f \in L^1_{loc}(\mathbb{R}^d)$ defines a unique distribution $u_f \in \mathcal{D}'(\mathbb{R}^d)$ by

$$\langle u_f, \phi\rangle = \int_{\mathbb{R}^d} f(\mathbf{x})\phi(\mathbf{x})d\mathbf{x} \qquad \forall \phi \in \mathcal{D}(\mathbb{R}^d).$$

In our notation we identify f with u_f.

Finally, in this subsection we state a result that is essentially the same as Lemma 5.1.1 on page 135 of [72], except we extend it from $d = 1$ to $d \in \mathbb{N}$. The proof is almost exactly the same for $d > 1$ and we present it in Appendix A.1. We will use this result later for proving Theorem 3.22.

Lemma 3.3. *Let $u \in \mathcal{D}'(\mathbb{R}^d)$ and let $K \subset \mathbb{R}^d$ be bounded. Then there exists a $n \in \mathbb{N}$ and a constant C_n such that*

$$|\langle u, \phi \rangle| \leq C_n \sum_{|\alpha| \leq n} \max_{\mathbf{x} \in K} |D^\alpha \phi(\mathbf{x})|$$

for all $\phi \in \mathcal{D}(\mathbb{R})$ with $\operatorname{supp} \phi \subset K$.

3.1.3 The Space $H^s(\mathbb{R}^d)$ for $s \in \mathbb{R}$

In this subsection we define the Sobolev space $H^s(\mathbb{R}^d)$ for $s \in \mathbb{R}$ via the Fourier Transform of temperate distributions. We begin by defining the Schwartz space of rapidly decreasing $C^\infty(\mathbb{R})$ functions. The definition is similar to the definition of $\mathcal{D}(\mathbb{R}^d)$.

Definition 3.4. Define the Schwartz space of rapidly decreasing C^∞ functions on $\mathbb{R}^d$ by

$$\mathcal{S}(\mathbb{R}^d) := \left\{ \phi \in C^\infty(\mathbb{R}^d) : \max_{\mathbf{x} \in \mathbb{R}^d} |\mathbf{x}^\alpha D^\beta \phi(\mathbf{x})| < \infty \text{ for all multi-indices } \alpha, \beta \right\}$$

Convergence in $\mathcal{S}(\mathbb{R}^d)$ is defined as follows: Let $\{\phi_n\}_{n=1}^\infty \subset \mathcal{S}(\mathbb{R}^d)$ be a sequence of functions in $\mathcal{S}(\mathbb{R}^d)$ and let $\phi \in \mathcal{S}(\mathbb{R}^d)$. We say that $\{\phi_n\}_{n=1}^\infty$ *converges to ϕ in $\mathcal{S}(\mathbb{R}^d)$* and write $\phi_n \xrightarrow{\mathcal{S}} \phi$ as $n \to \infty$ if

$$\max_{\mathbf{x} \in \mathbb{R}^d} |\mathbf{x}^\alpha D^\beta (\phi_n(\mathbf{x}) - \phi(\mathbf{x}))| \to 0 \qquad \text{as } n \to \infty,$$

for all multi-indices α, β.

We now define the space of temperate distributions in terms of functionals on $\mathcal{S}(\mathbb{R}^d)$.

Definition 3.5. A linear functional $u : \mathcal{S}(\mathbb{R}^d) \to \mathbb{R}$ is a temperate distribution on $\mathbb{R}^d$ if

$$\phi_n \xrightarrow{\mathcal{S}} \phi \qquad \Longrightarrow \qquad \langle u, \phi_n \rangle \to \langle u, \phi \rangle$$

for any $\phi_n, \phi \in \mathcal{S}(\mathbb{R}^d)$. The space of all temperate distributions on $\mathbb{R}^d$ is denoted by $\mathcal{S}'(\mathbb{R}^d)$.

Now we define the Fourier Transform for $u \in \mathcal{S}'(\mathbb{R}^d)$. If $u \in L^1(\mathbb{R}^d)$ then the Fourier transform of u is given by

$$\widehat{u}(\boldsymbol{\xi}) = \int_{\mathbb{R}^d} u(\mathbf{x})\, \mathrm{e}^{-i2\pi \boldsymbol{\xi} \cdot \mathbf{x}}\, d\mathbf{x}$$

for $\boldsymbol{\xi} \in \mathbb{R}^d$. For $u \in \mathcal{S}'(\mathbb{R}^d)$ the Fourier Transform of u is defined by

$$\langle \widehat{u}, \phi \rangle = \langle u, \widehat{\phi} \rangle \qquad \forall \phi \in \mathcal{S}(\mathbb{R}^d).$$

We can now define the space $H^s(\mathbb{R}^d)$ for $s \in \mathbb{R}$ as

$$H^s(\mathbb{R}^d) = \{u \in \mathcal{S}'(\mathbb{R}^d) : \|u\|_{H^s(\mathbb{R}^d)} < \infty\}$$

where

$$\|u\|_{H^s(\mathbb{R}^d)} = \left(\int_{\mathbb{R}^d} (1 + |\mathbf{k}|^2)^s |\widehat{u}(\mathbf{k})|^2 d\mathbf{k}\right)^{\frac{1}{2}}.$$

It follows from Plancheral's Theorem that $L^2(\mathbb{R}^d) = H^0(\mathbb{R}^d)$.

3.1.4 The Space $H^s(\Omega)$ for $s \in \mathbb{R}$

Now we define the Sobolev Space $H^s(\Omega)$ for $s \in \mathbb{R}$ and open, bounded $\Omega \subset \mathbb{R}^d$. It is defined as

$$H^s(\Omega) = \{u \in \mathcal{D}'(\Omega) : u = U|_\Omega \text{ for some } U \in H^s(\mathbb{R}^d)\}$$

with norm

$$\|u\|_{H^s(\Omega)} = \inf_{\substack{U \in H^s(\mathbb{R}^d) \\ U|_\Omega = u}} \|U\|_{H^s(\mathbb{R}^d)}.$$

We also define $H_0^s(\Omega)$ by

$$H_0^s(\Omega) = \text{closure of } \mathcal{D}(\Omega) \text{ in } H^s(\Omega).$$

3.1.5 The Standard Mollifier

In this subsection we define the standard mollifier for smoothing functions. We also present some of the basic properties of mollified functions. References for mollifiers include page 629 of [21] and page 36 of [2].

Definition 3.6. The standard mollifier $J \in C^\infty(\mathbb{R}^d)$ is defined by

$$J(\mathbf{x}) := \begin{cases} C \exp\left(\frac{1}{|\mathbf{x}|^2 - 1}\right) & |\mathbf{x}| < 1 \\ 0 & |\mathbf{x}| \geq 1, \end{cases}$$

where C is a constant chosen so that $\int_{\mathbb{R}^d} J(\mathbf{x}) d\mathbf{x} = 1$.

For $\epsilon > 0$ we also define $J_\epsilon(\mathbf{x}) := \epsilon^{-d} J(\epsilon^{-1}\mathbf{x})$. J_ϵ also has the property that $\int_{\mathbb{R}^d} J_\epsilon(\mathbf{x}) d\mathbf{x} = 1$.

Using $J_\epsilon(\mathbf{x})$ we can define a mollified function in the following way.

Definition 3.7. For $f \in L^1_{loc}(\mathbb{R}^d)$ and $\epsilon > 0$ we can define a mollified f by

$$f^{(\epsilon)}(\mathbf{x}) := J_\epsilon * f(\mathbf{x}) = \int_{B(0,\epsilon)} J_\epsilon(\mathbf{y}) f(\mathbf{x} - \mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^d} J_\epsilon(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y}$$

where $B(0,\epsilon) = \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}| < \epsilon\}$.

A mollified function has the following properties that are given in Theorem 6 on page 630 of [21].

Theorem 3.8. *If $f \in L^1_{loc}(\mathbb{R}^d)$ then*

1. *$f^{(\epsilon)} \in C^\infty(\mathbb{R}^d)$ for all $\epsilon > 0$.*
2. *$f^{(\epsilon)} \to f$ almost everywhere as $\epsilon \to 0$.*
3. *If $1 \leq p < \infty$ and $f \in L^p_{loc}(\mathbb{R}^d)$, then $f^{(\epsilon)} \to f$ in $L^p_{loc}(\mathbb{R}^d)$ as $\epsilon \to 0$.*

3.1.6 Estimating Series with Integrals

In this subsection we present a lemma that will allow us to estimate a series or partial series with an integral.

Lemma 3.9. *Let $p, q \in \mathbb{Z}$ with $p < q$, denote $I = [p, q] \subset \mathbb{R}$, and let $f \in C(I)$. Suppose that f is monotonically decreasing on I and $f(x) \geq 0$ for all $x \in I$. Then*

$$\sum_{n=p+1}^{q} f(n) \leq \int_I f(x)dx.$$

Conversely, if f is monotonically increasing on I then

$$\sum_{n=p}^{q-1} f(n) \leq \int_I f(x)dx.$$

Proof. We first consider the case when f is monotonically decreasing. Divide I into $(q-p)$ intervals of length 1, $I_j = [p+j-1, p+j]$ for $j = 1, \ldots, q-p$. Since f is monotonically decreasing $f(p+j) \leq f(x)$ for all $x \in I_j$ and $f(p+j) \leq \int_{I_j} f(x)dx$. Therefore,

$$\sum_{n=p+1}^{q} f(n) = \sum_{j=1}^{q-p} f(p+j) \leq \sum_{j=1}^{q-p} \int_{I_j} f(x)dx = \int_I f(x)dx$$

The proof for f monotonically increasing is similar. □

Lemma 3.9 can be extended to infinite series by taking the limit as $q \to \infty$ (in the case when f is monotonically decreasing).

3.2 Periodic Functions

In this section we develop the theory of periodic functions and their representation using plane waves (or Fourier basis functions). We begin by defining periodic functions

and the Fourier Series of functions in $L^1_{loc}(\mathbb{R}^d)$. We then define Periodic Sobolev Spaces and we present a few embedding theorems for Periodic Sobolev Spaces. Next, we relate Periodic Sobolev Spaces back to usual Sobolev spaces by presenting a result about equivalent norms. Following that, we define two finite dimensional periodic function spaces in terms of the span of a finite number of plane waves. We also define the Fourier representation and nodal representation of functions in these finite dimensional spaces. We then describe the Discrete Fourier Transform and its implementation, the Fast Fourier Transform, as a way of swapping between these two representations of functions in our finite dimensional spaces. Finally, we define projections onto our finite dimensional function spaces and we quote some estimates for the difference between a function and its projection.

While most of the results in this section are needed for developing theoretical error bounds for our problem, the Fast Fourier Transform is the crucial ingredient for an efficient implementation of our method.

Throughout this section we will endeavour to present results that are general for functions defined on $\mathbb{R}^d$ for $d \in \mathbb{N}$, although we only need the results for $d \in \{1, 2\}$ in this thesis.

Before we continue, we must define what a periodic function is. We do this by first defining a Bravais lattice. We will also need the definition of the reciprocal lattice. A good reference for lattice definitions is [3].

Definition 3.10. Let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_d$ be d linearly independent vectors in $\mathbb{R}^d$. A d-dimensional *Bravais lattice* $\mathbf{R}$ is the set of points

$$\mathbf{R} := \left\{ \mathbf{r} \in \mathbb{R}^d : \mathbf{r} = \sum_{j=1}^{d} n_j \mathbf{a}_j,\ n_j \in \mathbb{Z} \right\}$$

The vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_d$ are called *primitive lattice vectors*. The *Wigner-Seitz primitive cell* $\mathbf{W}$ is defined as the set of points closer to the origin than any other lattice point,

$$\mathbf{W} := \left\{ \mathbf{x} \in \mathbb{R}^d : |\mathbf{x}| < \min_{\mathbf{r} \in \mathbf{R} \setminus \{\mathbf{0}\}} |\mathbf{x} + \mathbf{r}| \right\}$$

We note that the primitive lattice vectors are not unique for a given Bravais lattice. There are also other ways of chosing the primitive cell but we will use the Wigner-Seitz primitive cell in this thesis. Another name for the Wigner-Seitz primitive cell is the Voronoi cell.

In addition to defining the Bravais lattice we also need to define the corresponding reciprocal lattice and the 1st Brillouin zone.

Definition 3.11. Let $\mathbf{R}$ be a Bravais lattice in $\mathbb{R}^d$. The *reciprocal lattice* $\mathbf{R}_c$ is also a

Bravais lattice and it is defined by

$$\mathbf{R}_c := \{\mathbf{k} \in \mathbb{R}^d : \mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}} = 1,\ \forall \mathbf{r} \in \mathbf{R}\}$$

The Wigner-Seitz primitive cell of the reciprocal lattice is called the *1st Brillouin zone*.

Definition 3.12. A function $f : \mathbb{R}^d \to \mathbb{C}$ is periodic if, for some Bravais lattice $\mathbf{R}$ in $\mathbb{R}^d$,

$$f(\mathbf{x}) = f(\mathbf{x} + \mathbf{r}) \qquad \forall \mathbf{r} \in \mathbf{R}, \mathbf{x} \in \mathbb{R}^d.$$

We denote the period cell of f with Ω, and it is defined as the Wigner-Seitz primitive cell of $\mathbf{R}$.

Conversely, given a periodic function with period cell Ω, we have implicitly defined a Bravais lattice, with a primitive cell that is equal to Ω, as well as a reciprocal lattice that has a 1st Brillouin Zone.

With this definition of periodicity in mind, it is clear that any function defined on Ω, where Ω is the primitive cell of a lattice, can be extended to a periodic function on all of $\mathbb{R}^d$ in the sense of Definition 3.12.

Given a Bravais lattice we can also define periodic function spaces. For example,

$$\begin{aligned} L^1_p &= \{f \in L^1_{loc}(\mathbb{R}^d) : f \text{ is periodic with period cell } \Omega\} \\ L^2_p &= \{f \in L^2_{loc}(\mathbb{R}^d) : f \text{ is periodic with period cell } \Omega\} \\ C_p(\Omega) &= \{f \in C(\mathbb{R}^d) : f \text{ is periodic with period cell } \Omega\} \\ C^\infty_p &= \{f \in C^\infty(\mathbb{R}^d) : f \text{ is periodic with period cell } \Omega\}. \end{aligned}$$

We will often write C_p instead of $C_p(\Omega)$ when it is obvious that C_p is a function space and not a constant. We equip $C_p(\Omega)$ with the uniform norm

$$\|u\|_\infty = \max_{\mathbf{x}\in\mathbb{R}^d} |u(\mathbf{x})|.$$

For the rest of this thesis we will restrict ourselves to the most basic Bravais lattice in $\mathbb{R}^d$, namely $\mathbb{Z}^d$. The Weigner-Seitz primitive cell is $\Omega := (-\frac{1}{2}, \frac{1}{2})^d$ and the 1st Brillouin zone is $B := (-\pi, \pi)^d$. Although we make this restriction, all of the results could be extended to more general lattices by using an appropriate change of variables that maps the general lattice back onto $\mathbb{Z}^d$.

If a function is not periodic in every coordinate direction then we will specify this. For example, a function defined on $\mathbb{R}^2$ that is only periodic in the x-direction will be called x-periodic.

3.2.1 Fourier Series

In this subsection we define the Fourier Series for periodic functions defined on $\mathbb{R}^d$. We will also define Fourier coefficients. The definition of Fourier coefficients will be used extensively throughout the rest of this thesis. Here is a definition of the Fourier Series.

Definition 3.13. The *Fourier Series* of $f \in L_p^1$ is defined as

$$\sum_{\mathbf{g}\in\mathbb{Z}^d} [f]_{\mathbf{g}}\, \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}$$

where $[f]_{\mathbf{g}}$ is the *Fourier coefficient* of f with index $\mathbf{g}$ and is defined by

$$[f]_{\mathbf{g}} := \int_{\Omega} f(\mathbf{x})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x}.$$

Throughout the rest of this thesis we will use square brackets, $[\cdot]_{\mathbf{g}}$, to denote the Fourier coefficient of a function with index $\mathbf{g}$.

The following result is a special case of a theorem in Chapter 1 of [16].

Theorem 3.14. *For the case $d = 1$: If a periodic function f is piecewise continuous with a finite number of maxima and minima on Ω, then*

$$\lim_{N\to\infty} \sum_{k=-N}^{N} [f]_k\, \mathrm{e}^{i2\pi kx} = \lim_{\epsilon\searrow 0} \frac{f(x+\epsilon) + f(x-\epsilon)}{2}, \qquad x \in \mathbb{R}.$$

There are other results that we could quote with respect to the convergence of the Fourier Series in $\mathbb{R}$. In particular, in 1D a piecewise continuous function with a finite number of maxima and minima on Ω (that is absolutely continuous on intervals of continuity) is a special case of a function with bounded variation for which Theorem 3.14 also holds. This result is known as Jordan's Criterion according to [16].

We use Theorem 3.14 to identify all piecewise continuous functions with finitely many maxima and minima on Ω with their Fourier Series *everywhere* in $\mathbb{R}$. The result is that we can write

$$f(x) = \sum_{k\in\mathbb{Z}} [f]_k\, \mathrm{e}^{i2\pi kx} \qquad \forall x \in \mathbb{R}.$$

For Fourier Series in $\mathbb{R}^d$ for $d > 1$ there are greater restrictions on f to obtain pointwise convergence. According to [80, Theorem 1.7 on page 248] the trigonometric polynomials are dense in $C_p(\Omega)$ for arbitrary d (with norm $\|\cdot\|_\infty$), and it follows from [80, Corollary 1.8] that if $f \in C_p(\Omega)$ and $\sum_{\mathbf{g}} |[f]_{\mathbf{g}}| < \infty$ then its Fourier Series converges everywhere to f. We are interested in the pointwise convergence of the Fourier Series for discontinuous functions. For $d = 2$, [60, Theorem 1] implies that if $f \in L_p^1$ and f has bounded variation then the Fourier Series of f converges everywhere. The piecewise continuous functions

in $\mathbb{R}^2$ that we will define in Section 3.3 satisfy the definition of bounded variation in [60] and we can at least be sure that the Fourier Series converges pointwise everywhere to something. However, for the definition of the projection Q_n in Subsection 3.2.5 to be well-defined for discontinuous functions we would like to know what that something is. The most useful result in the literature that we could find to help us resolve this problem is in [67]. In [67] the authors describe a function space that includes some discontinuous functions for which we get pointwise convergence of the Fourier Series for $d \geq 1$. We will (as briefly as possible) present their result for $d = 2$. For notational convenience we only consider convergence at the point $\mathbf{x} = 0$. Define an alternative period cell $\Omega' = [0, 1]^2$, the interval $I = (0, 1/2)$, let Ω'_0 denote the interior of Ω' and let f^* define the following function,

$$f^*(x, y) = f(x, y) + f(-x, y) + f(x, -y) + f(-x, -y).$$

Now we define the set of functions $\mathcal{F}$, where $f \in \mathcal{F}$ if $f \in L^1_p$ and there exists $g \in L^1_p$ such that $f = g$ on Ω'_0, $g_1, g_2 \in L^1(I)$ and $g_{12} \in L^1(I^2)$ where

$$\begin{aligned} g_1(t) &= \frac{g^*(t, 0) - g^*(0)}{t} \\ g_2(t) &= \frac{g^*(0, t) - g^*(0)}{t} \\ g_{12}(s, t) &= \frac{g^*(s, t) - g^*(s, 0) - g^*(0, t) + g^*(0)}{st}. \end{aligned}$$

[67, Theorem 4.2] then states that if $f \in \mathcal{F}$ and that, for some open ball B centred at 0, f^* is continuous on $\Omega'_0 \cap B$ and has a continuous extension to $\partial\Omega' \cap \overline{B}$, then the Fourier Series of f at 0 converges to

$$\lim_{N \to \infty} \sum_{|n_i| \leq N} [f]_{\mathbf{n}} \, \mathrm{e}^{i2\pi\mathbf{n}\cdot 0} = \lim_{\epsilon \to 0} \frac{f^*(\epsilon, \epsilon)}{4}.$$

Now we must ask: In more practical terms, what functions are in $\mathcal{F}$? It is immediate that if $f \in L^1_p$ and is smooth in a neighbourhood of 0, then $f \in \mathcal{F}$ and the Fourier Series of f converges to f at 0. In this thesis we will mostly be interested in piecewise constant functions so we restrict the rest of this discussion to this type of function and we consider the case when f is discontinuous at 0. Let B be an open ball centred at 0 with radius $\delta > 0$, let $m \in \mathbb{R}$ and consider functions $f \in L^1_p$ such that

$$f(\mathbf{x}) = \begin{cases} f_1 & x_2 > m x_1 \\ \frac{1}{2}(f_1 + f_2) & x_2 = m x_1 \\ f_2 & x_2 < m x_1 \end{cases} \quad \text{for all } \mathbf{x} \in B$$

or

$$f(\mathbf{x}) = \begin{cases} f_1 & x_1 < 0 \\ \frac{1}{2}(f_1 + f_2) & x_1 = 0 \\ f_2 & x_1 > 0 \end{cases} \text{ for all } \mathbf{x} \in B.$$

It is possible to check that with f defined in this way we have $f \in \mathcal{F}$ and so the Fourier Series of f at 0 converges to $\frac{1}{2}(f_1 + f_2)$. The final discontinuous function that we consider has the form,

$$f(\mathbf{x}) = \begin{cases} f_1 & x_1 < 0 \text{ or } x_2 < 0 \\ f_2 & x_1 > 0 \text{ and } x_2 > 0 \\ \frac{1}{2}(f_1 + f_2) & x_2 > 0 \text{ and } x_1 = 0 \\ \frac{1}{2}(f_1 + f_2) & x_1 > 0 \text{ and } x_2 = 0 \\ \frac{3}{4}f_1 + \frac{1}{4}f_2 & \mathbf{x} = 0 \end{cases} \text{ for all } \mathbf{x} \in B$$

It can be shown that this function also belongs to $\mathcal{F}$ and its Fourier Series converges at 0.

Other functions with this type of corner where the interfaces are aligned with the coordinate axes are admissible in $\mathcal{F}$. Unfortunately, functions with corners or curved interfaces are generally not in $\mathcal{F}$ and we do not know what the Fourier Series converges to at these points.

Before we move onto Periodic Sobolev Spaces, let us state the following lemma. It states that the Fourier coefficients of functions in C_p^∞ decay superalgebraically.

Lemma 3.15. *Let $\phi \in C_p^\infty$. Then for any $r \in \mathbb{N}$ there exists a constant C_r such that $|[\phi]_{\mathbf{n}}| \leq C_r |\mathbf{n}|^{-r}$ for all $\mathbf{0} \neq \mathbf{n} \in \mathbb{Z}^d$.*

Proof. The proof of this result can be obtained by applying integration by parts to the formula for $[\phi]_{\mathbf{n}}$ in Definition 3.13. □

3.2.2 Periodic Sobolev Spaces

In this subsection we define Periodic Sobolev Spaces H_p^s for $s \in \mathbb{R}$ and include some results about these spaces that will be useful in the rest of this thesis. We first define Periodic Sobolev Spaces on $\mathbb{R}^d$ for $d \in \mathbb{N}$ before restricting ourselves to $d \in \{1, 2\}$ for particular results.

All of this subsection is based on the theory presented in [72] where the definition of Periodic Sobolev Spaces for $d = 1$ is presented as well as results for $d \in \{1, 2\}$. Periodic distributions for $d = 2$ are used in [72] but they are not explicitly defined. In this subsection we extend the definitions in [72] to $d \in \mathbb{N}$. All of the results for $d \in \{1, 2\}$ are quoted from [72], except for Theorem 3.29.

Other references for Periodic Sobolev Spaces include [18] and [52]. In [18], Sobolev spaces are defined on a C^∞ smooth closed curve in the complex plane whereas in [52], Sobolev spaces are defined on a C^∞ class boundary of a bounded, open set in $\mathbb{R}^d$. By using an appropriate parameterization of the curve or boundary it can be shown that Periodic Sobolev Spaces are special cases of these Sobolev spaces. To our knowledge, [72] is the most detailed reference on Periodic Sobolev Spaces.

We begin by defining periodic distributions and we extend the definition of Fourier coefficients in Definition 3.13 to periodic distributions. We then use the definition of Fourier coefficients for periodic distributions to define Periodic Sobolev Spaces. We finish the subsection by presenting some embedding results for Periodic Sobolev Spaces, interpolation results for Periodic Sobolev Spaces, estimates for periodic distributions multiplied by continuous functions and a result that shows the equivalence of the periodic Sobolev space norms to usual Sobolev space norms.

First, we define what it means to say that a distribution is periodic.

Definition 3.16. A distribution $u \in \mathcal{D}'(\mathbb{R}^d)$ is periodic if

$$\langle u, \tau_{\mathbf{n}}\phi\rangle = \langle u, \phi\rangle \qquad \forall \phi \in \mathcal{D}(\mathbb{R}^d), \mathbf{n} \in \mathbb{Z}^d$$

where $(\tau_{\mathbf{n}}\phi)(\mathbf{x}) = \phi(\mathbf{x}+\mathbf{n})$ for all $\mathbf{x} \in \mathbb{R}^d$. We denote the set of all periodic distributions by $\mathcal{D}'_p(\mathbb{R}^d)$.

Now that we have defined periodic distributions, we extend our definition of Fourier coefficients to include the Fourier coefficients of periodic distributions. We do this in the same way as in [72] except we extend their theory to $\mathcal{D}'_p(\mathbb{R}^d)$ with $d > 1$. We begin by presenting the following result which defines a partition of unity for $\mathbb{R}^d$.

Lemma 3.17. *There exists a function $\theta \in \mathcal{D}(\mathbb{R}^d)$ such that $0 \le \theta(\mathbf{x}) \le 1$ for all $\mathbf{x} \in \mathbb{R}^d$, $\operatorname{supp}\theta \subset \widetilde{\Omega} = (-\frac{3}{2}, \frac{3}{2})^d$, and*

$$\sum_{\mathbf{n}\in\mathbb{Z}^d} \tau_{\mathbf{n}}\theta(\mathbf{x}) = \sum_{\mathbf{n}\in\mathbb{Z}^d} \theta(\mathbf{x}+\mathbf{n}) = 1 \qquad \forall \mathbf{x} \in \mathbb{R}^d.$$

Moreover, if $V \subset\subset \Omega = (-\frac{1}{2}, \frac{1}{2})^d$ then we can define θ such that $\theta(\mathbf{x}) = 1$ for all $\mathbf{x} \in V$.

Proof. On page 137 of [72] we can find a result that says there exists a function $\theta_1 \in \mathcal{D}(\mathbb{R})$ such that $\sum_{n\in\mathbb{Z}} \theta_1(x+n) = 1$ for all $x \in \mathbb{R}$. In [72] they prove their result by constructing an example that satisfies $\sum_{n\in\mathbb{Z}} \theta_1(x+n) = 1$ for all $x \in \mathbb{R}$. Their example also satisfies $0 \le \theta_1(x) \le 1$ for all $x \in \mathbb{R}$ and $\operatorname{supp}\theta_1 \subset (-\frac{3}{2}, \frac{3}{2})$.

We use θ_1 to construct θ. Define

$$\theta(\mathbf{x}) = \prod_{i=1}^{d} \theta_1(x_i) \qquad \forall \mathbf{x} \in \mathbb{R}^d$$

Then

$$\sum_{\mathbf{n}\in\mathbb{Z}^d} \theta(\mathbf{x}+\mathbf{n}) = \sum_{\mathbf{n}\in\mathbb{Z}^d} \prod_{i=1}^{d} \theta_1(x_i+n_i) = \prod_{i=1}^{d} \sum_{n_i\in\mathbb{Z}} \theta_1(x_i+n_i) = 1 \qquad \forall \mathbf{x}\in\mathbb{R}^d$$

It is obvious that $0 \le \theta(\mathbf{x}) \le 1$ for all $\mathbf{x}\in\mathbb{R}$ and $\operatorname{supp}\theta \subset \widetilde{\Omega}$.

For the second part of Lemma 3.17 we construct θ_1 and θ. Define

$$\epsilon := \inf_{\substack{\mathbf{x}\in V \\ \mathbf{y}\in\partial\Omega}} |\mathbf{x}-\mathbf{y}| \qquad \text{and} \qquad 1_\Omega(\mathbf{x}) := \begin{cases} 1 & \mathbf{x}\in\Omega \\ 0 & \mathbf{x}\notin\Omega \end{cases}.$$

Set $\theta_1(x) = J_\epsilon * 1_\Omega(x)$ (see Subsection 3.1.5) and $\theta(\mathbf{x}) = \prod_{i=1}^d \theta_1(x_i)$. To complete the proof it is enough to show that $\theta_1(x_i) = 1$ for $i = 1,\dots,d$ and all $\mathbf{x}\in V$ and $\sum_{n\in\mathbb{Z}} \theta_1(x+n) = 1$ for all $x\in\mathbb{R}$.

Let $\mathbf{x}\in V$. Then by the definition of ϵ we have that $1_\Omega(x_i-y) = 1$ for all $y\in B(0,\epsilon)$ and so

$$\theta_1(x_i) = \int_{B(0,\epsilon)} J_\epsilon(y) 1_\Omega(x_i-y)dy = \int_{B(0,\epsilon)} J_\epsilon(y)dy = 1$$

We also get, using the fact that $\sum_{n\in\mathbb{Z}} 1_\Omega(x+n-y) = 1$ for almost every $x,y\in\mathbb{R}$,

$$\begin{aligned} \sum_{n\in\mathbb{Z}} \theta_1(x+n) &= \sum_{n\in\mathbb{Z}} \int_{\mathbb{R}} J_\epsilon(y) 1_\Omega(x+n-y)dy \\ &= \int_{\mathbb{R}} J_\epsilon(y) \left(\sum_{n\in\mathbb{Z}} 1_\Omega(x+n-y) \right) dy \\ &= \int_{\mathbb{R}} J_\epsilon(y)dy \\ &= 1 \qquad\qquad \forall x\in\mathbb{R}. \end{aligned}$$

□

See Figure 3-1 for a plot of a θ that satisfies Lemma 3.17 in 1D. Now, using a θ defined as in Lemma 3.17 we define the Fourier coefficients for periodic distributions.

Definition 3.18. Let $u\in\mathcal{D}'_p(\mathbb{R}^d)$ be a periodic distribution and let $\theta\in\mathcal{D}(\mathbb{R}^d)$ be defined as in Lemma 3.17. Then the Fourier coefficient of u with index $\mathbf{g}\in\mathbb{Z}^d$ is defined by

$$[u]_{\mathbf{g}} = \langle u, \psi \rangle$$

where $\psi(\mathbf{x}) = \theta(\mathbf{x})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}} \in \mathcal{D}(\mathbb{R}^d)$.

From this definition it appears that the Fourier coefficient of $u\in\mathcal{D}'_p(\mathbb{R}^d)$ depends on the choice of θ. We will show in Lemma 3.20 that this is not the case.

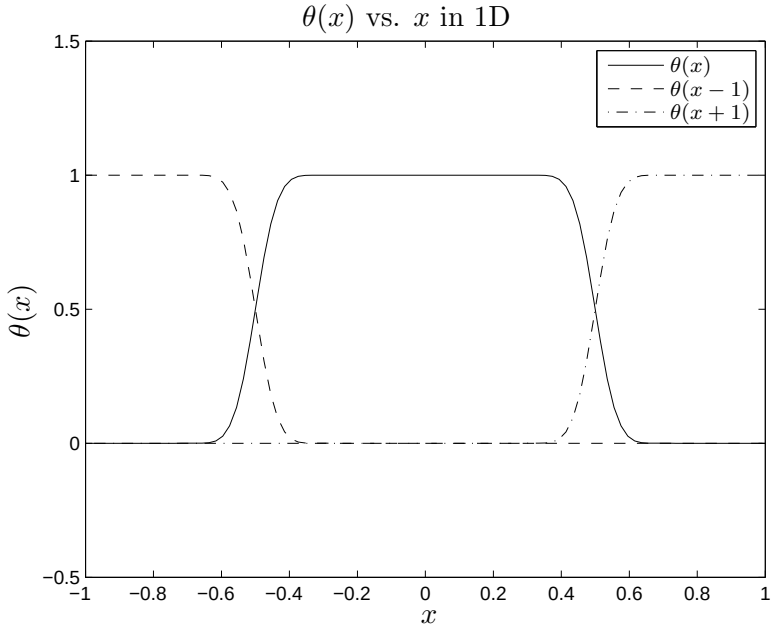

Figure 3-1: Here is an example of a possible $\theta(x)$ in 1D from Lemma 3.17. For $|x| \in [\frac{1}{4}, \frac{3}{4}]$, $\theta_1(x) = \frac{f(a-|x|)}{f(a-|x|)+f(|x|-b)}$ where $f(x) = \mathrm{e}^{-1/x}$, $a = \frac{1}{4}$ and $b = \frac{3}{4}$.

Instead of defining periodic distributions in terms of functionals on the space of test functions with compact support, sometimes it is more convenient to define periodic distributions as functionals on a set of test functions that are periodic.

Definition 3.19. We define the space of periodic test functions on $\mathbb{R}^d$ as

$$\mathcal{D}_p(\mathbb{R}^d) = C_p^\infty.$$

Convergence in $\mathcal{D}_p(\mathbb{R}^d)$ is defined as follows: Let $\{\phi_n\}_{n=1}^\infty \subset \mathcal{D}_p(\mathbb{R}^d)$ be a set of test functions and let $\phi \in \mathcal{D}_p(\mathbb{R}^d)$. We say ϕ_n *converges to* ϕ *in* $\mathcal{D}_p(\mathbb{R}^d)$ and write $\phi_n \xrightarrow{\mathcal{D}_p} \phi$ as $n \to \infty$ if

$$\|D^\alpha(\phi_n - \phi)\|_\infty \to 0$$

as $n \to \infty$, for any multi-index α. We also define a new duality for $\mathcal{D}_p'(\mathbb{R}^d)$ and $\mathcal{D}_p(\mathbb{R}^d)$ by

$$\langle u, \phi\rangle_p := \langle u, \theta\phi\rangle \qquad \forall u \in \mathcal{D}_p'(\mathbb{R}^d), \phi \in \mathcal{D}_p(\mathbb{R}^d)$$

where θ satisfies Lemma 3.17. Finally, we define convergence of $u_n, u \in \mathcal{D}_p'(\mathbb{R}^d)$,

$$u_n \to u \text{ in } \mathcal{D}_p(\mathbb{R}^d) \qquad \text{if} \qquad \langle u_n, \phi\rangle_p \to \langle u, \phi\rangle_p \qquad \forall \phi \in \mathcal{D}_p(\mathbb{R}^d).$$

Lemma 3.20. *For $u \in \mathcal{D}_p'(\mathbb{R}^d)$, $\mathbf{g} \in \mathbb{Z}^d$ and $\phi \in \mathcal{D}_p(\mathbb{R}^d)$ the Fourier coefficient $[u]_\mathbf{g}$*

and the dual product $\langle u, \phi\rangle_p$ are independent from the choice of θ satisfying Lemma 3.17.

Proof. If θ and $\widetilde{\theta}$ both satisfy Lemma 3.17, then

$$\begin{aligned}
\langle u, \widetilde{\theta}\phi\rangle &= \left\langle u, \sum_{\mathbf{n}\in\mathbb{Z}^d} (\tau_{\mathbf{n}}\theta)\widetilde{\theta}\phi \right\rangle && \text{by Lemma 3.17}\\
&= \sum_{\mathbf{n}\in\mathbb{Z}^d} \langle u, (\tau_{\mathbf{n}}\theta)\widetilde{\theta}\phi\rangle && \text{by linearity}\\
&= \sum_{\mathbf{n}\in\mathbb{Z}^d} \langle u, \tau_{-\mathbf{n}}[(\tau_{\mathbf{n}}\theta)\widetilde{\theta}\phi]\rangle && \text{by Definition 3.16}\\
&= \sum_{\mathbf{n}\in\mathbb{Z}^d} \langle u, \theta(\tau_{-\mathbf{n}}\widetilde{\theta})\phi\rangle && \text{since } \phi \text{ is periodic}\\
&= \langle u, \theta\phi\rangle && \text{by linearity and Lemma 3.17.}
\end{aligned}$$

Therefore, $\langle u, \phi\rangle_p$ is independent from the choice of θ that satisfies Lemma 3.17.

The proof for $[u]_{\mathbf{g}}$ independent of θ is obtained by choosing $\phi(\mathbf{x}) = \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}$ in the argument above. □

We extend Lemma 5.2.1 on page 139 of [72] to get the following result for $d > 1$. It shows that convergence in $\mathcal{D}'_p(\mathbb{R}^d)$ is equivalent to convergence in $\mathcal{D}'(\mathbb{R}^d)$. The proof is almost exactly the same as the proof given in [72] for the $d = 1$ case and we omit it.

Lemma 3.21. *For $u_n, u \in \mathcal{D}'_p(\mathbb{R}^d)$ the following statements are equivalent*

1. *$u_n \to u$ in $\mathcal{D}'_p(\mathbb{R}^d)$, i.e. $\langle u_n, \phi\rangle_p \to \langle u, \phi\rangle_p$ for all $\phi \in \mathcal{D}_p(\mathbb{R}^d)$;*
2. *$u_n \to u$ in $\mathcal{D}'(\mathbb{R}^d)$, i.e. $\langle u_n, \psi\rangle \to \langle u, \psi\rangle$ for all $\psi \in \mathcal{D}(\mathbb{R}^d)$.*

Recall that we have a defined Fourier coefficients of periodic distributions in Definition 3.18. However, we cannot yet be sure that we can write

$$u(\mathbf{x}) = \sum_{\mathbf{n}\in\mathbb{Z}^d} [u]_{\mathbf{n}}\, \mathrm{e}^{i2\pi\mathbf{n}\cdot\mathbf{x}} \qquad \text{in } \mathcal{D}'_p(\mathbb{R}^d). \tag{3.1}$$

The next theorem addresses this problem as well as proving some basic properties of periodic distributions and periodic test functions. It is an obvious extension of Theorem 5.2.1 on page 140 of [72].

Theorem 3.22. *Let $u \in \mathcal{D}'_p(\mathbb{R}^d)$ and $\phi \in \mathcal{D}_p(\mathbb{R}^d)$. Then*

1. *There exists a $k \in \mathbb{N}$ and consant C_k such that $|[u]_{\mathbf{n}}| \le C_k|\mathbf{n}|^k$ for all $\mathbf{0} \neq \mathbf{n} \in \mathbb{Z}^d$,*
2. *$\langle u, \phi\rangle_p = \sum_{n\in\mathbb{Z}^d} [u]_{\mathbf{n}}[\phi]_{-\mathbf{n}}$,*
3. *$\sum_{|\mathbf{n}|\le N} [u]_{\mathbf{n}}\, \mathrm{e}^{i2\pi\mathbf{n}\cdot\mathbf{x}} \to u(x)$ in $\mathcal{D}'_p(\mathbb{R}^d)$ as $N \to \infty$.*

Proof. We prove Part 1 using Definition 3.18 and Lemma 3.3. With $\mathbf{0} \neq \mathbf{n} \in \mathbb{Z}^d$,

$$\begin{aligned}
|[u]_{\mathbf{n}}| &= |\langle u, \psi \rangle| && \text{by Def. 3.18 with } \psi(\mathbf{x}) = \theta(\mathbf{x})\, \mathrm{e}^{-i2\pi \mathbf{n}\cdot\mathbf{x}} \\
&\leq C_k \sum_{|\alpha| \leq k} \max_{\mathbf{x} \in \operatorname{supp} \theta} |D^\alpha \psi(\mathbf{x})| && \text{with } k \in \mathbb{N} \text{ from Theorem 3.3} \\
&\leq C_k' |\mathbf{n}|^k && \text{since } \psi(\mathbf{x}) = \theta(\mathbf{x})\, \mathrm{e}^{-i2\pi \mathbf{n}\cdot\mathbf{x}}
\end{aligned}$$

Part 2. Since ϕ is continuous we can write it in terms of its Fourier Series. With θ defined according to Lemma 3.17 we get

$$\begin{aligned}
\langle u, \phi \rangle_p &= \langle u, \theta\phi \rangle && \text{by Definition 3.19} \\
&= \left\langle u(\mathbf{x}), \theta(\mathbf{x}) \sum_{\mathbf{n} \in \mathbb{Z}^d} [\phi]_{\mathbf{n}}\, \mathrm{e}^{i2\pi \mathbf{n}\cdot\mathbf{x}} \right\rangle \\
&= \sum_{\mathbf{n} \in \mathbb{Z}^d} [\phi]_{\mathbf{n}} \langle u(\mathbf{x}), \theta(\mathbf{x})\, \mathrm{e}^{i2\pi \mathbf{n}\cdot\mathbf{x}} \rangle \\
&= \sum_{\mathbf{n} \in \mathbb{Z}^d} [\phi]_{\mathbf{n}} [u]_{-\mathbf{n}} && \text{by Definition 3.18} \\
&= \sum_{\mathbf{m} \in \mathbb{Z}^d} [u]_{\mathbf{m}} [\phi]_{-\mathbf{m}}.
\end{aligned}$$

Part 3. Finally, we use Parts 1 and 2 and Lemma 3.15 to prove Part 3. Let $\phi \in \mathcal{D}_p(\mathbb{R}^d)$. Then there exists a constant C_s such that

$$\begin{aligned}
\left\langle \sum_{|\mathbf{n}| \leq N} [u]_{\mathbf{n}}\, \mathrm{e}^{i2\pi \mathbf{n}\cdot\mathbf{x}} - u(x), \phi \right\rangle_p &= \sum_{|\mathbf{n}| > N} [u]_{\mathbf{n}} [\phi]_{-\mathbf{n}} && \text{by Part 2} \\
&\leq C_s \sum_{|\mathbf{n}| > N} |\mathbf{n}|^{-s} && \forall s \in \mathbb{N} \text{ by Part 1 and Lem. 3.15}
\end{aligned}$$

which converges to 0 as $N \to \infty$. □

Part 3 of Theorem 3.22 ensures that we can identify $u \in \mathcal{D}_p'(\mathbb{R}^d)$ with its Fourier Series as in (3.1).

Now we define Periodic Sobolev Spaces in terms of the decay of these Fourier coefficients as the magnitude of the index of the Fourier coefficients increases.

Definition 3.23. We define the following Periodic Sobolev Space and norm for $s \in \mathbb{R}$

$$H_p^s = \{u \in \mathcal{D}_p'(\mathbb{R}^d) : \|u\|_{H_p^s} < \infty\}$$

where

$$\|u\|_{H_p^s} = \left(\sum_{\mathbf{n} \in \mathbb{Z}^d} |\mathbf{n}|_\star^{2s} |[u]_{\mathbf{n}}|^2 \right)^{\frac{1}{2}} \qquad \text{and} \qquad |\mathbf{n}|_\star = \begin{cases} 1 & \mathbf{n} = \mathbf{0} \\ |\mathbf{n}| & \mathbf{n} \neq \mathbf{0} \end{cases}$$

H_p^s is complete with respect to this norm and it is a Hilbert space with inner product

$$(u,v)_{H_p^s} = \sum_{\mathbf{n}\in\mathbb{Z}^d} |\mathbf{n}|_\star^{2s}[u]_{\mathbf{n}}\overline{[v]_{\mathbf{n}}} \qquad \text{for } u,v \in H_p^s.$$

We may write (by expanding u and v in terms of their Fourier Series and then integrating)

$$(u,v)_{H_p^0} = \int_\Omega u(\mathbf{x})\overline{v(\mathbf{x})}d\mathbf{x} \qquad \text{for } u,v \in L_p^2 \tag{3.2}$$

and so $H_p^0 = L_p^2$.

For $s \in \mathbb{R}$, $u \in H_p^s$ and $v \in H_p^{-s}$ we can write (again, by expanding u and v in terms of their Fourier Series and using the Cauchy-Schwarz inequality)

$$|(u,v)_{H_p^0}| = \left|\int_\Omega u\overline{v}d\mathbf{x}\right| = \left|\sum_{\mathbf{n}\in\mathbb{Z}^d} (|\mathbf{n}|_\star^s[u]_{\mathbf{n}})(|\mathbf{n}|_\star^{-s}\overline{[v]_{\mathbf{n}}})\right| \leq \|u\|_{H_p^s}\|v\|_{H_p^{-s}} \tag{3.3}$$

We can also extend $\langle\cdot,\cdot\rangle_p$ defined on $\mathcal{D}_p'(\mathbb{R}^d)\times\mathcal{D}_p(\mathbb{R}^d)$ to $H_p^s \times H_p^{-s}$ for $s \in \mathbb{R}$. We get (using same arguement as in Part 2 of Theorem 3.22)

$$\langle u,v\rangle_p = \sum_{\mathbf{n}\in\mathbb{Z}^d} [u]_{\mathbf{n}}[v]_{-\mathbf{n}}$$

and similarly to (3.3) we can write

$$|\langle u,v\rangle_p| \leq \|u\|_{H_p^s}\|v\|_{H_p^{-s}} \tag{3.4}$$

for $u \in H_p^s$ and $v \in H_p^{-s}$. Furthermore, for all $u \in H_p^s$ there exists a $v \in H_p^{-s}$ with $\|v\|_{H_p^{-s}} = 1$ such that $\|u\|_{H_p^s} = \langle u,v\rangle_p$ (for $u \neq 0$ take v with Fourier coefficients $[v]_{\mathbf{n}} = |\mathbf{n}|_\star^s\overline{[u]_{-\mathbf{n}}}/\|u\|_{H_p^s}$, $\mathbf{n}\in\mathbb{Z}^d$). From this we can write

$$\|u\|_{H_p^s} = \max_{v\in H_p^{-s}} \frac{|\langle u,v\rangle_p|}{\|v\|_{H_p^{-s}}} \qquad \forall u \in H_p^s. \tag{3.5}$$

From the definition of the norm $\|\cdot\|_{H_p^s}$, it is obvious that we have $H_p^t \subset H_p^s$ for $s \leq t$. When $s < t$ we find that the embedding is compact. The following result is an exercise on page 143 of [72].

Lemma 3.24. *If $s < t$ then*

$$H_p^t \subset\subset H_p^s.$$

Proof. As we have already mentioned, it is obvious from the definition of the norm that $H_p^t \subset H_p^s$. To show that the embedding is compact we must show that the inclusion operator $I : H_p^t \to H_p^s$ is compact.

For $N \in \mathbb{N}$ define an operator $\mathrm{P}_N : H_p^t \to H_p^s$ by

$$\mathrm{P}_N\, u(\mathbf{x}) = \sum_{|\mathbf{n}|\le N} [u]_{\mathbf{n}}\, \mathrm{e}^{i2\pi\mathbf{n}\cdot\mathbf{x}} \qquad \forall \mathbf{x} \in \mathbb{R}^d$$

for all $u \in H_p^t$. P_N is bounded and has finite rank. Therefore, P_N is a compact operator.

Now we show that $\mathrm{P}_N \to I$ in the operator norm as $N \to \infty$. Let $u \in H_p^t$ and $N \in \mathbb{N}$. Then

$$\begin{aligned}
\|(I - \mathrm{P}_N)u\|_{H_p^s} &= \|u - \mathrm{P}_N\, u\|_{H_p^s} \\
&= \left(\sum_{|\mathbf{n}|>N} |\mathbf{n}|^{2s} |[u]_{\mathbf{n}}|^2 \right)^{\frac{1}{2}} \\
&= \left(\sum_{|\mathbf{n}|>N} |\mathbf{n}|^{2s-2t} |\mathbf{n}|^{2t} |[u]_{\mathbf{n}}|^2 \right)^{\frac{1}{2}} \\
&\le \left(N^{2s-2t}\right)^{1/2} \left(\sum_{|\mathbf{n}|>N} |\mathbf{n}|^{2t} |[u]_{\mathbf{n}}|^2 \right)^{\frac{1}{2}} \\
&\le N^{s-t} \|u\|_{H_p^t}.
\end{aligned}$$

Therefore, $\|I - \mathrm{P}_n\|_{\mathcal{L}(H_p^t, H_p^s)} \le N^{s-t} \to 0$ as $N \to \infty$ since $s < t$.

The result then follows from the fact that a limit of a sequence of compact operators with finite rank must also be compact. □

Now we present two interpolation results. The first result is an extension of Lemma 5.12.2 on page 162 of [72] for $d > 1$ while the second result is an exercise from [72]. The proof of Lemma 3.25, although it is an extension to what is in [72], is exactly the same as the one given in [72]. We will present a proof of Lemma 3.26. Both results rely on a result called *The Three Lines Theorem* (also given in [72]). We include the details of The Three Lines Theorem in the proof of Lemma 3.26.

Lemma 3.25. *Let A be an operator such that $A \in \mathcal{L}(H_p^{s_1}, H_p^{t_1})$ and $A \in \mathcal{L}(H_p^{s_2}, H_p^{t_2})$ for $s_1, s_2, t_1, t_2 \in \mathbb{R}$ with $s_1 \le s_2$ and $t_1 \le t_2$. Then, for $\tau \in [0, 1]$,*

$$\|A\|_{\mathcal{L}(H_p^{\tau s_1 + (1-\tau)s_2}, H_p^{\tau t_1 + (1-\tau)t_2})} \le \|A\|^{\tau}_{\mathcal{L}(H_p^{s_1}, H_p^{t_1})} \|A\|^{1-\tau}_{\mathcal{L}(H_p^{s_2}, H_p^{t_2})}$$

Lemma 3.26. *Let $s, t \in \mathbb{R}$ with $s \le t$, $u \in H_p^t$ and $\tau \in [0, 1]$. Then*

$$\|u\|_{H_p^{\tau s + (1-\tau)t}} \le \|u\|^{\tau}_{H_p^s} \|u\|^{1-\tau}_{H_p^t}$$

Proof. This proof uses *The Three Lines Theorem* (Lemma 5.12.1 in [72]). It is stated as follows: Let $F(z)$ be a continuous function in the closed strip $z = x + iy$, $a \le x \le b$, $y \in \mathbb{R}$. Assume that $F(z)$ is analytic and bounded in the open strip $a < x < b$, $y \in \mathbb{R}$. With $M(x) := \sup_{y\in\mathbb{R}} |F(x + iy)|$, we get

$$M(x) \le M(a)^{\frac{b-x}{b-a}} M(b)^{\frac{x-a}{b-a}} \qquad a \le x \le b. \tag{3.6}$$

In this proof we will also need to define the operator, $\Lambda^z : H^\mu \to H^{\mu-\mathrm{Re}z}$, for $z \in \mathbb{C}$ and $\mu \in \mathbb{R}$, by

$$(\Lambda^z u)(\mathbf{t}) = \sum_{\mathbf{n}\in\mathbb{Z}^d} |\mathbf{n}|_\star^z [u]_\mathbf{n} \, \mathrm{e}^{i2\pi\mathbf{n}\cdot\mathbf{t}} \qquad \mathbf{t} \in \mathbb{R}.$$

Since $||\mathbf{n}|_\star^z| = |\mathbf{n}|_\star^{\mathrm{Re}z}$, we get

$$\|\Lambda^z u\|_{H_p^\mu} = \|\Lambda^{\mathrm{Re}z} u\|_{H_p^\mu} = \|u\|_{H_p^{\mu+\mathrm{Re}z}} \qquad \forall u \in H_p^\mu, z \in \mathbb{C}, \mu \in \mathbb{R}. \tag{3.7}$$

For $u \in H_p^t$, $v \in H_p^0$ and $z \in \mathbb{C}$ with $s \le \mathrm{Re}z \le t$, let us define

$$F(z) := \langle \Lambda^z u, v\rangle_p = \sum_{\mathbf{n}\in\mathbb{Z}^d} |\mathbf{n}|_\star^z [u]_\mathbf{n} [v]_\mathbf{n}.$$

Since $|\mathbf{n}|_\star^z$ is analytic with respect to z for all $\mathbf{n} \in \mathbb{Z}^d$, $F(z)$ is analytic. Moreover, $F(z)$ is bounded (see (3.4)). Therefore, we can apply (3.6) with $a = s$, $b = t$ and $x = \tau s + (1 - \tau)t$ to get

$$\begin{aligned}
|\langle \Lambda^{\tau s+(1-\tau)t} u, v\rangle_p| &= |F(\tau s + (1-\tau)t)| \\
&\le \sup_{y\in\mathbb{R}} |F(\tau s + (1-\tau)t + iy)| \\
&\le \left(\sup_{y\in\mathbb{R}} |F(s + iy)|\right)^\tau \left(\sup_{y\in\mathbb{R}} |F(t + iy)|\right)^{1-\tau} && \text{by (3.6)} \\
&\le \left(\sup_{y\in\mathbb{R}} \|\Lambda^{s+iy} u\|_{H_p^0} \|v\|_{H_p^0}\right)^\tau \left(\sup_{y\in\mathbb{R}} \|\Lambda^{t+iy} u\|_{H_p^0} \|v\|_{H_p^0}\right)^{1-\tau} && \text{by (3.4)} \\
&= \|u\|_{H_p^s}^\tau \|v\|_{H_p^0}^\tau \|u\|_{H_p^t}^{1-\tau} \|v\|_{H_p^0}^{1-\tau} && \text{by (3.7)} \\
&= \|u\|_{H_p^s}^\tau \|u\|_{H_p^t}^{1-\tau} \|v\|_{H_p^0} && (3.8)
\end{aligned}$$

Now we use (3.7), (3.5) and (3.8) to get

$$\|u\|_{H_p^{\tau s+(1-\tau)t}} = \|\Lambda^{\tau s+(1-\tau)t} u\|_{H_p^0} = \sup_{v\in H_p^0} \frac{|\langle \Lambda^{\tau s+(1-\tau)t} u, v\rangle_p|}{\|v\|_{H_p^0}} \le \|u\|_{H_p^s}^\tau \|u\|_{H_p^t}^{1-\tau}$$

□

For the remainder of this subsection we will restrict ourselves to distributions on $\mathbb{R}^d$ with $d \in \{1, 2\}$.

We now state another embedding theorem for Periodic Sobolev Spaces.

Theorem 3.27. *1. Let $d = 1$ and $s > \frac{1}{2}$. Then $u \in H^s_p(\Omega)$ is continuous and*

$$\|u\|_\infty \leq C_s \|u\|_{H^s_p(\Omega)}$$

where $C_s = (\sum_{n\in\mathbb{Z}} |n|_\star^{-2s})^{1/2}$.

2. Let $d = 2$ and $s > 1$. Then $u \in H^s_p(\Omega)$ is continuous and

$$\|u\|_\infty \leq C_s \|u\|_{H^s_p(\Omega)}$$

where $C_s = (\sum_{\mathbf{n}\in\mathbb{Z}^2} |\mathbf{n}|_\star^{-2s})^{1/2}$.

Proof. Both of these results are Sobolev Embedding Theorems. The statement and proof of part 1 is Lemma 5.3.2 on page 142 of [72] while the statement of part 2 is exercise 8.5.4 on page 254 of [72]. The proof of Part 2 is very similar to the proof of part 1 and we present it now.

Let $u_N(\mathbf{x}) = \sum_{|\mathbf{n}|\leq N} [u]_\mathbf{n} \, \mathrm{e}^{i2\pi\mathbf{n}\cdot\mathbf{x}}$. Then

$$\begin{aligned}\|u_N\|_\infty &\leq \sum_{|\mathbf{n}|\leq N} |[u]_\mathbf{n}| \leq \sum_{|\mathbf{n}|\leq N} |[u]_\mathbf{n}||\mathbf{n}|_\star^s |\mathbf{n}|_\star^{-s} \leq \left(\sum_{|\mathbf{n}|\leq N} |[u]_\mathbf{n}|^2 |\mathbf{n}|_\star^{2s}\right)^{\frac{1}{2}} \left(\sum_{|\mathbf{n}|\leq N} |\mathbf{n}|_\star^{-2s}\right)^{\frac{1}{2}} \\ &\leq C_s \|u\|_{H^s_p}\end{aligned}$$

and so

$$\|u_N - u_M\|_\infty \leq C_s \|u_N - u_M\|_{H^s_p} \to 0, \qquad N, M \to \infty$$

The result follows from the fact that $C_p(\Omega)$ is complete with respect to $\|\cdot\|_\infty$. □

Finally, in this subsection we state some estimates for a distribution from a Periodic Sobolev Space multiplied by sufficiently smooth periodic function.

Theorem 3.28. *1. With $d = 1$, for $s \in \mathbb{R}$, $t > 1/2$, $a \in H_p^{\max(|s|,t)}$ and $u \in H^s_p$ then there exist constants C_s and C_t such that*

$$\|au\|_{H^s_p} \leq C_s \|a\|_{H_p^{|s|}} \|u\|_{H^s_p} \qquad \text{for } |s| > \tfrac{1}{2}$$

and

$$\|au\|_{H^s_p} \leq C_t \|a\|_{H^t_p} \|u\|_{H^s_p} \qquad \text{for } |s| \leq \tfrac{1}{2}$$

2. *With $d = 2$, for $s \in \mathbb{R}$, $t > 1$, $a \in H_p^{\max(|s|,t)}$ and $u \in H_p^s$ then there exist constants C_s and C_t such that*

$$\|au\|_{H_p^s} \le C_s \|a\|_{H_p^{|s|}} \|u\|_{H_p^s} \qquad \text{for } |s| > 1$$

and

$$\|au\|_{H_p^s} \le C_t \|a\|_{H_p^t} \|u\|_{H_p^s} \qquad \text{for } |s| \le 1$$

Proof. Part 1 is Lemma 5.13.1 on page 163 of [72], except that the statement of the Lemma in [72] requires that $a \in C_p^\infty$. This is too conservative and the proof given in [72] goes through for $a \in H_p^{\max(|s|,t)}$ as we have stated. Part 2 is not in [72]. The proof is very similar to the proof of Part 1 and we present it now.

We have

$$\begin{aligned} a(\mathbf{x})u(\mathbf{x}) &= \sum_{\mathbf{m}\in\mathbb{Z}^2} [a]_{\mathbf{m}} \, \mathrm{e}^{i2\pi\mathbf{m}\cdot\mathbf{x}} \sum_{\mathbf{n}\in\mathbb{Z}^2} [u]_{\mathbf{n}} \, \mathrm{e}^{i2\pi\mathbf{n}\cdot\mathbf{x}} \\ &= \sum_{\mathbf{m},\mathbf{n}\in\mathbb{Z}^2} [a]_{\mathbf{m}}[u]_{\mathbf{n}} \, \mathrm{e}^{i2\pi(\mathbf{m}+\mathbf{n})\cdot\mathbf{x}} \\ &= \sum_{\mathbf{k}\in\mathbb{Z}^2} \left(\sum_{\mathbf{n}\in\mathbb{Z}^2} [a]_{\mathbf{k}-\mathbf{n}}[u]_{\mathbf{n}} \right) \mathrm{e}^{i2\pi\mathbf{k}\cdot\mathbf{x}} \end{aligned}$$

and so we may write

$$\|au\|_{H_p^s} \le \left\{ \sum_{\mathbf{k}\in\mathbb{Z}^2} \left(\sum_{\mathbf{n}\in\mathbb{Z}^2} |\mathbf{k}|_\star^s |[a]_{\mathbf{k}-\mathbf{n}}||[u]_{\mathbf{n}}| \right)^2 \right\}^{\frac{1}{2}} \qquad (s \in \mathbb{R}) \tag{3.9}$$

Now we split into different cases according to s.

Case $s > 1$. Using $|\mathbf{k}|_\star^s \le 2^s(|\mathbf{k}-\mathbf{n}|_\star^s + |\mathbf{n}|_\star^s)$ and (3.9) we get

$$\begin{aligned} \|au\|_{H_p^s} &= \left\{ \sum_{\mathbf{k}\in\mathbb{Z}^2} \left(|\mathbf{k}|_\star |[au]_{\mathbf{k}}| \right)^2 \right\}^{\frac{1}{2}} \\ &= \left\{ \sum_{\mathbf{k}\in\mathbb{Z}^2} \left(|\mathbf{k}|_\star \left| \sum_{\mathbf{n}\in\mathbb{Z}^2} [a]_{\mathbf{k}-\mathbf{n}}[u]_{\mathbf{n}} \right| \right)^2 \right\}^{\frac{1}{2}} \\ &\le 2^s \left\{ \sum_{\mathbf{k}\in\mathbb{Z}^2} \left(\sum_{\mathbf{n}\in\mathbb{Z}^2} |\mathbf{k}-\mathbf{n}|_\star^s |[a]_{\mathbf{k}-\mathbf{n}}||[u]_{\mathbf{n}}| + \sum_{\mathbf{n}\in\mathbb{Z}^2} |[a]_{\mathbf{k}-\mathbf{n}}||\mathbf{n}|_\star^s |[u]_{\mathbf{n}}| \right)^2 \right\}^{\frac{1}{2}} \\ &= 2^s \|bv + dw\|_{H_p^0} \le 2^s (\|bv\|_{H_p^0} + \|dw\|_{H_p^0}) \end{aligned} \tag{3.10}$$

where the functions b, v, d, w are defined by their Fourier coefficients,

$$[b]_{\mathbf{k}} = |\mathbf{k}|_\star^s |[a]_{\mathbf{k}}| \qquad [v]_{\mathbf{n}} = |[u]_{\mathbf{n}}|$$
$$[d]_{\mathbf{k}} = |[a]_{\mathbf{k}}| \qquad [w]_{\mathbf{n}} = |\mathbf{n}|_\star^s |[u]_{\mathbf{n}}|$$

for $\mathbf{k}, \mathbf{n} \in \mathbb{Z}^2$. We have $\|a\|_{H_p^s} = \|b\|_{H_p^0} = \|d\|_{H_p^s}$ and $\|u\|_{H_p^s} = \|v\|_{H_p^s} = \|w\|_{H_p^0}$. By (3.2) and Theorem 3.27 we get

$$\|bv\|_{H_p^0} = \left(\int_\Omega |b(\mathbf{x})v(\mathbf{x})|^2 d\mathbf{x} \right)^{\frac{1}{2}} \leq \|b\|_{H_p^0} \|v\|_\infty \leq C_s \|b\|_{H_p^0} \|v\|_{H_p^s} = C_s \|a\|_{H_p^s} \|u\|_{H_p^s}$$
$$\|dw\|_{H_p^0} = \left(\int_\Omega |d(\mathbf{x})w(\mathbf{x})|^2 d\mathbf{x} \right)^{\frac{1}{2}} \leq \|d\|_\infty \|w\|_{H_p^0} \leq C_s \|d\|_{H_p^s} \|w\|_{H_p^0} = C_s \|a\|_{H_p^s} \|u\|_{H_p^s}$$

The result follows from (3.10) and is

$$\|au\|_{H_p^s} \leq 2^{s+1} C_s \|a\|_{H_p^s} \|u\|_{H_p^s} \qquad \text{for } s > 1. \tag{3.11}$$

Case $s = 0$. This result follows from (3.2) and Theorem 3.27 using the fact that $t > 1$ and $a \in H_p^t$,

$$\|au\|_{H_p^0} = \left(\int_\Omega |a(\mathbf{x})u(\mathbf{x})|^2 d\mathbf{x} \right)^{\frac{1}{2}} \leq \|a\|_\infty \|u\|_{H_p^0} \leq C_s \|a\|_{H_p^t} \|u\|_{H_p^0} \tag{3.12}$$

Case $0 < s \leq 1$. Now we apply the interpolation result in Lemma 3.25 where A is the multiplication operator defined by $Au = au$. The inequality (3.11) implies that $A \in \mathcal{L}(H_p^t, H_p^t)$ for $t > 1$ while (3.12) implies that $A \in \mathcal{L}(H_p^0, H_p^0)$. Applying Lemma 3.25 yields $A \in \mathcal{L}(H_p^{(1-\tau)t}, H_p^{(1-\tau)t})$ for $0 \leq \tau \leq 1$ and

$$\|A\|_{\mathcal{L}(H_p^{(1-\tau)t}, H_p^{(1-\tau)t})} \leq (C_s \|a\|_{H_p^t})^\tau (2^{s+1} C_s \|a\|_{H_p^t})^{1-\tau} = 2^{(s+1)(1-\tau)} C_s \|a\|_{H_p^t}.$$

The result is then

$$\|a\|_{H_p^{(1-\tau)t}} \leq 2^{(s+1)(1-\tau)} C_s \|a\|_{H_p^t} \|u\|_{H_p^{(1-\tau)t}} \qquad \text{for } t > 1, 0 \leq \tau \leq 1.$$

Case $s < 0$. This case is proved using a duality argument that is the same as in the $d = 1$ proof in [72]. □

Now we present a result that shows how $\|\cdot\|_{H_p^s}$ is related to the usual Sobolev space norms.

Theorem 3.29. *For $s \geq 0$ and with θ defined as in Lemma 3.17,*

$$\|u\|_{H_p^s} \simeq \|u\|_{H^s(\Omega)} \simeq \|\theta u\|_{H^s(\mathbb{R}^d)} \qquad \forall u \in H_p^s. \tag{3.13}$$

Proof. Let $s \geq 0$ and suppose $u \in H_p^s$. The result $\|u\|_{H_p^s} \simeq \|u\|_{H^s(\Omega)}$ is from Chapter 5 of [18]. However, in [18], the norm $\|\cdot\|_{H^s(\Omega)}$ is defined as the Slobodeckiĭ norm, whereas we have defined $\|\cdot\|_{H^s(\Omega)}$ in terms of the Fourier tranform (see Subsection 3.1.4). A result proving when these two norms are equivalent is given in Theorem 3.18 of [54].

The second result, $\|u\|_{H^s(\Omega)} \simeq \|\theta u\|_{H^s(\mathbb{R}^d)}$, follows from the following simple argument. Define $\theta \in \mathcal{D}(\mathbb{R}^d)$ and $\widetilde{\Omega}$ as in Lemma 3.17. Define

$$\overline{\theta}(\mathbf{x}) = \sum_{\substack{\mathbf{n}\in\mathbb{Z}^d \\ |n_i|\leq 1}} \theta(\mathbf{x}+\mathbf{n}) \qquad \forall \mathbf{x} \in \mathbb{R}^d.$$

Then $\overline{\theta}(\mathbf{x}) = 1$ for all $\mathbf{x} \in \Omega$ and by the Definition of $\|\cdot\|_{H^s(\Omega)}$,

$$\begin{aligned}\|u\|_{H^s(\Omega)} \leq \|\overline{\theta}u\|_{H^s(\mathbb{R}^d)} &\leq \sum_{|n_i|\leq 1} \|\theta(\mathbf{x}+\mathbf{n})u(\mathbf{x})\|_{H^s(\mathbb{R}^d)} \\ &= \sum_{|n_i|\leq 1} \|\theta(\mathbf{x}+\mathbf{n})u(\mathbf{x}+\mathbf{n})\|_{H^s(\mathbb{R}^d)} = 3^d\|\theta u\|_{H^s(\mathbb{R}^d)}.\end{aligned}$$

Conversely, there is a constant C (that depends on θ and s) such that

$$\|\theta u\|_{H^s(\mathbb{R}^d)} = \|\theta u\|_{H^s(\widetilde{\Omega})} \leq C\|u\|_{H^s(\widetilde{\Omega})} = 3^d C\|u\|_{H^s(\Omega)}.$$

□

3.2.3 Trigonometric Function Spaces

In this section we define two types of finite dimensional function spaces which consist of functions that are in the span of a finite number of plane waves (or Fourier basis functions).

First, we define some notation. For $d \in \mathbb{N}$ (we only need $d \in \{1, 2\}$) and $n \in \mathbb{N}$, denote

$$\begin{aligned}\mathbb{Z}^d_{n,\mathrm{o}} &= \Big\{\mathbf{n} \in \mathbb{Z}^d : |\mathbf{n}| \leq n\Big\} \\ \mathbb{Z}^d_{n,\square} &= \Big\{\mathbf{n} \in \mathbb{Z}^d : -\frac{n}{2} \leq n_i < \frac{n}{2}, i = 1, \ldots, d\Big\}\end{aligned}$$

where $|\cdot|$ denotes the usual Euclidean norm of a vector. For $d = 1$, $\mathbb{Z}^1_{n,\mathrm{o}} = \mathbb{Z}^1_{2n+1,\square}$. Using these definitions we define

$$\begin{aligned}\mathcal{S}_n^{(d)} &= \mathrm{span}\{\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} : \mathbf{g} \in \mathbb{Z}^d_{n,\mathrm{o}}\} \\ \mathcal{T}_n^{(d)} &= \mathrm{span}\{\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} : \mathbf{g} \in \mathbb{Z}^d_{n,\square}\}\end{aligned}$$

When it is obvious we will omit the superscript and just write $\mathcal{S}_n$ or $\mathcal{T}_n$. For $d = 1$, we get $\mathcal{T}_{2n+1} = \mathcal{S}_n$, $\dim \mathcal{S}_n = 2n + 1$ and $\dim \mathcal{T}_n = n$. For $d = 2$, $\dim \mathcal{S}_n = \mathcal{O}(n^2)$

and $\dim \mathcal{T}_n = n^2$. The set $\{\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} : \mathbf{g} \in \mathbb{Z}^d_{n,\mathrm{o}}\}$ is an orthogonal basis for $\mathcal{S}_n$ where orthogonality is with respect to the $L^2(\Omega)$ inner product. Similarly, $\{\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} : \mathbf{g} \in \mathbb{Z}^d_{n,\square}\}$ is an orthogonal basis for $\mathcal{T}^{(d)}_n$. We will call each of these bases a *Fourier basis* and each member of the basis set will be a *Fourier basis function*. Since we have a basis, every function $f \in \mathcal{S}^{(d)}_n$ can be expanded uniquely as a linear combination of the Fourier basis functions and we can write

$$f(\mathbf{x}) = \sum_{\mathbf{g}\in\mathbb{Z}^d_{n,\mathrm{o}}} c_{\mathbf{g}}\, \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}\,. \tag{3.14}$$

where $c_{\mathbf{g}} = [f]_{\mathbf{g}}$ are constants. We will refer to this expansion of $f \in \mathcal{S}^{(d)}_n$ as the *Fourier representation* of f. An alternative way of expressing this is to recognize that if we have a vector (for $d = 1$) or a matrix (for $d = 2$) of Fourier coefficients $c_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^d_{n,\mathrm{o}}$ then we have uniquely defined a function $f(\mathbf{x}) \in \mathcal{S}^{(d)}_n$ according to (3.14). We will also refer to a vector or matrix of Fourier coefficients as the *Fourier representation* of a function.

We can also define a Fourier representation of $f \in \mathcal{T}^{(d)}_n$ in a similar way.

3.2.4 Discrete and Fast Fourier Transforms

In this subsection we will consider functions in $\mathcal{T}^{(d)}_n$. We will show that as well as having a Fourier representation of $f \in \mathcal{T}^{(d)}_n$, there is also a *nodal representation* of f (we do not define a nodal representation for functions in $\mathcal{S}^{(d)}_n$). We will then present the Discrete Fourier Transform (DFT) which is a transform for switching between these two representations. Finally, we discuss the Fast Fourier Transform (FFT) which is a very efficient algorithm for computing the DFT and its inverse.

Before we define the nodal representation of $f \in \mathcal{T}^{(d)}_n$ we must define the following function in $\mathcal{T}^{(1)}_n$. For $n \in \mathbb{N}$ and $k \in \mathbb{Z}^1_{n,\square}$,

$$\phi_{n,k}(x) = \frac{1}{n} \sum_{j\in\mathbb{Z}^1_{n,\square}} \mathrm{e}^{i2\pi j(x-k/n)} = \sum_{j\in\mathbb{Z}^1_{n,\square}} \left(\frac{1}{n}\,\mathrm{e}^{-i2\pi jk/n}\right) \mathrm{e}^{i2\pi jx}\,.$$

The function $\phi_{n,k}$ is a linear combination of the Fourier basis functions of $\mathcal{T}^{(1)}_n$ and it has the following property,

$$\phi_{n,k}(\tfrac{m}{n}) = \delta_{mk} \qquad \text{for } m \in \mathbb{Z}^1_{n,\square}.$$

The functions $\phi_{k,n}$ for different $k \in \mathbb{Z}_n$ are also orthogonal with respect to the $L^2(\Omega)$ inner product.

Using $\phi_{n,k}$ we define the *nodal representation* of $f \in \mathcal{T}_n^{(d)}$ as

$$f(\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{Z}_{n,\square}^d} d_{\mathbf{k}} \, \varphi_{n,\mathbf{k}}^{(d)}(\mathbf{x}) \tag{3.15}$$

where

$$d_{\mathbf{k}} = f(\tfrac{1}{n}\mathbf{k}) \qquad \text{and} \qquad \varphi_{n,\mathbf{k}}^{(d)}(\mathbf{x}) = \prod_{i=1}^{d} \phi_{n,k_i}(x_i).$$

We see that the coefficients $d_{\mathbf{k}}$ are the nodal values $f(\mathbf{x})$ where the nodes are a uniform grid with grid-spacing $\frac{1}{n}$ and it can be shown that the set $\{\varphi_{n,\mathbf{k}}^{(d)}(x) : \mathbf{k} \in \mathbb{Z}_{n,\square}^d\}$ is an orthogonal basis for $\mathcal{T}_n^{(d)}$. We call this basis of $\mathcal{T}_n^{(d)}$ the nodal basis and each member of the basis is called a nodal basis function. An alternative interpretation of the nodal representation is to recognize that if we know the values of a function in $\mathcal{T}_n^{(d)}$ at the nodes $\{\frac{1}{n}\mathbf{k} : \mathbf{k} \in \mathbb{Z}_{n,\square}^d\}$, then the function is uniquely determined. A vector (for $d = 1$) or a matrix (for $d = 2$) of nodal values, since it uniquely defines a function in $\mathcal{T}_n^{(d)}$, will also be referred to as the *nodal representation* of a function in $\mathcal{T}_n^{(d)}$.

We have now seen that we can represent a function $f \in \mathcal{T}_n^{(d)}$ using either the Fourier representation or nodal representation. We saw that we can store f as a vector or a matrix of either Fourier coefficients $\{c_{\mathbf{g}} = [f]_{\mathbf{g}} : \mathbf{g} \in \mathbb{Z}_{n,\square}^d\}$ or nodal values $\{d_{\mathbf{k}} = f(\frac{1}{n}\mathbf{k}) : \mathbf{k} \in \mathbb{Z}_{n,\square}^d\}$. The Discrete Fourier Transform (DFT) specifies the Fourier coefficients of f in terms of the nodal values of f and the Inverse Discrete Fourier Transform (IDFT) specifies the nodal values of f in terms of the Fourier coefficients of f. It is defined as follows.

$$c_{\mathbf{g}} = \frac{1}{n} \sum_{\mathbf{k} \in \mathbb{Z}_{n,\square}^d} d_{\mathbf{k}} \, \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{k}/n} \qquad \forall \mathbf{g} \in \mathbb{Z}_{n,\square}^d \tag{DFT}$$

$$d_{\mathbf{k}} = \sum_{\mathbf{g} \in \mathbb{Z}_{n,\square}^d} c_{\mathbf{g}} \, \mathrm{e}^{i2\pi \mathbf{g}\cdot\mathbf{k}/n} \qquad \forall \mathbf{k} \in \mathbb{Z}_{n,\square}^d. \tag{IDFT}$$

The Fast Fourier Transform (FFT) is an algorithm that is able to compute the Discrete Fourier Transform in $\mathcal{O}(n^d \log n)$ operations for any $n \in \mathbb{N}$. However, the performance of the FFT algorithm is the most efficient when $n = 2^k$ for $k \in \mathbb{N}$. The Fast Fourier Transform was first published in [10], although we use the implementation developed by [30].

We finish this subsection by fixing some notation for the case when $d = 2$. Consider a function $f \in \mathcal{T}_n^{(2)}$ where n is even. As per our discussion above f can be uniquely determined with either n^2 Fourier coefficients or n^2 nodal values. We store these values in $n \times n$ matrices X and $\widehat{X}$. Our convention is to store the nodal values in X and the Fourier coefficients in $\widehat{X}$. We also have a special indexing convention for these matrices.

Let $m = \frac{n}{2} + 1$. Then

$$X_{ij} = f\left(\tfrac{(i-m,j-m)}{n}\right)$$
$$\widehat{X}_{ij} = [f]_{(i-m,j-m)}$$

for $i, j = 1, \ldots, n$. We can now express the 2D FFT and inverse FFT as operators on matrices. We denote the 2D FFT by $\mathrm{fft}(\cdot)$ and the 2D inverse FFT by $\mathrm{ifft}(\cdot)$. For example, we get $\widehat{X} = \mathrm{fft}(X)$ and $X = \mathrm{ifft}(\widehat{X})$.

3.2.5 Orthogonal and Interpolation Projections

In this subsection we define projections from H_p^s onto $\mathcal{S}_n^{(d)}$ and $\mathcal{T}_n^{(d)}$ and we also derive some estimates for these projections. We will define the projections in a natural way that associates them with either the Fourier representation or nodal representation of a function in either $\mathcal{S}_n$ or $\mathcal{T}_n$.

We begin by defining the Orthogonal Projections, $\mathrm{P}_n^{(\mathcal{S})} : H_p^s \to \mathcal{S}_n^{(d)}$ and $\mathrm{P}_n^{(\mathcal{T})} : H_p^s \to \mathcal{T}_n^{(d)}$. For $s \in \mathbb{R}$, $u \in H_p^s$ and $n \in \mathbb{N}$, they are defined by

$$\mathrm{P}_n^{(\mathcal{S})}\, u(\mathbf{x}) = \sum_{\mathbf{g} \in \mathbb{Z}_{n,\mathrm{o}}^d} [u]_{\mathbf{g}}\, \mathrm{e}^{i2\pi \mathbf{g}\cdot\mathbf{x}}$$
$$\mathrm{P}_n^{(\mathcal{T})}\, u(\mathbf{x}) = \sum_{\mathbf{g} \in \mathbb{Z}_{n,\square}^d} [u]_{\mathbf{g}}\, \mathrm{e}^{i2\pi \mathbf{g}\cdot\mathbf{x}}$$

for all $\mathbf{x} \in \mathbb{R}^d$. We will now state some estimates for these two projections.

Lemma 3.30. *For $s, t \in \mathbb{R}$ with $s \le t$, $d \in \{1, 2\}$ and $n \in \mathbb{N}$, if $u \in H_p^t$ then*

$$\|u - \mathrm{P}_n^{(\mathcal{S})}\, u\|_{H_p^s} \le n^{s-t} \|u\|_{H_p^t} \tag{3.16}$$
$$\|u - \mathrm{P}_n^{(\mathcal{T})}\, u\|_{H_p^s} \le (\tfrac{n}{2})^{s-t} \|u\|_{H_p^t}. \tag{3.17}$$

Proof. The results in (3.16) and (3.17) for $d = 1$ are essentially the same since $\mathcal{S}_n = \mathcal{T}_{2n+1}$ in 1D and (3.17) for $d = 1$ is Theorem 8.2.1 on page 241 of [72].

The result in (3.17) for $d = 2$ is Lemma 8.5.1 on page 253 of [72] whereas (3.16) for $d = 2$ is not in [72]. We prove (3.16) for $d = 2$ now. The proof is very similar to the proof of the $d = 1$ result. For $s, t \in \mathbb{R}$, $s \le t$, $u \in H_p^t$ and $n \in \mathbb{N}$ we get

$$\begin{aligned}\|u - \mathrm{P}_n^{(\mathcal{S})}\, u\|_{H_p^s}^2 &= \sum_{\mathbf{n} \in \mathbb{Z}^2 \setminus \mathbb{Z}_{n,\mathrm{o}}^2} |\mathbf{n}|_\star^{2s} |[u]_{\mathbf{n}}|^2 = \sum_{|\mathbf{n}| > n} |\mathbf{n}|^{2(s-t)} |\mathbf{n}|^{2t} |[u]_{\mathbf{n}}|^2 \\ &\le n^{2(s-t)} \sum_{|\mathbf{n}| > n} ||\mathbf{n}|^{2t} |[u]_{\mathbf{n}}|^2 \le n^{2(s-t)} \|u\|_{H_p^t}^2.\end{aligned}$$

□

Now we move onto defining the Interpolation Projection, $\mathrm{Q}_n : C_p(\Omega) \to \mathcal{T}_n^{(d)}$ (there is no Q projection onto $\mathcal{S}_n^{(d)}$). It is naturally associated with the nodal representation of a trigonometric function. For a continuous periodic function u defined on $\mathbb{R}^d$ and $n \in \mathbb{N}$ we define

$$\mathrm{Q}_n\, u \in \mathcal{T}_n^{(d)} \quad \text{such that} \quad (\mathrm{Q}_n\, u)(\tfrac{1}{n}\mathbf{k}) = u(\tfrac{1}{n}\mathbf{k}) \quad \forall \mathbf{k} \in \mathbb{Z}^d_{n,\square}.$$

From our definition of the nodal representation of functions in $\mathcal{T}_n^{(d)}$ we know that this uniquely defines a projection onto $\mathcal{T}_n^{(d)}$.

If u is discontinuous then $\mathrm{Q}_n\, u$ may not be well-defined but we can extend the definition of Q_n to distributions that have a convergent Fourier Series. In this case Q_n is defined by nodal values that are given by the Fourier Series of u,

$$\mathrm{Q}_n\, u(\tfrac{1}{n}\mathbf{k}) = \sum_{\mathbf{g}\in\mathbb{Z}^d} [u]_{\mathbf{g}}\, \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{k}/n} \qquad \forall \mathbf{k} \in \mathbb{Z}^d_{n,\square}.$$

By the definition of this projection we automatically obtain the nodal representation of $\mathrm{Q}_n\, u \in \mathcal{T}_n^{(d)}$. We know that there also exists a Fourier representation of $\mathrm{Q}_n\, u$. The following Lemma gives us the Fourier coefficients of $\mathrm{Q}_n\, u$. It is explicitly stated in Lemma 8.3.1 on page 242 of [72] for the case when $d = 1$ and u is continuous. It is also implicitly used on page 251 of [72] for the case when $d = 2$. Here we state a more general result than that stated in [72] in the sense that we let $d \in \mathbb{N}$ and we let u be possibly discontinuous.

Lemma 3.31. *Let $d \in \mathbb{N}$ and let u be a periodic function on $\mathbb{R}^d$ with a convergent Fourier Series. Then*

$$[\mathrm{Q}_n\, u]_{\mathbf{g}} = \sum_{\mathbf{k}\in\mathbb{Z}^d} [u]_{\mathbf{g}+n\mathbf{k}} \qquad \forall \mathbf{g} \in \mathbb{Z}^d_{n,\square}$$

Proof. This proof is very similar to the proof of Lemma 8.3.1 on page 242 in [72]. We have $\mathrm{Q}_n\, v = v$ for all $v \in \mathcal{T}_n^{(d)}$. In particular, we have $\mathrm{Q}_n\, \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} = \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}$ for all $\mathbf{g} \in \mathbb{Z}^d_{n,\square}$. We also have, for $\mathbf{g} \in \mathbb{Z}^d_{n,\square}$ and $\mathbf{k} \in \mathbb{Z}^d$,

$$\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} = \mathrm{e}^{i2\pi(\mathbf{g}+n\mathbf{k})\cdot\mathbf{x}}$$

at $\mathbf{x} = \frac{1}{n}\mathbf{m}$ for $\mathbf{m} \in \mathbb{Z}^d_{n,\square}$ since $\mathrm{e}^{i2\pi n\mathbf{k}\cdot\mathbf{m}/n} = 1$. That is, $\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}$ and $\mathrm{e}^{i2\pi(\mathbf{g}+n\mathbf{k})\cdot\mathbf{x}}$ have the same nodal values. Therefore,

$$\mathrm{Q}_n\, \mathrm{e}^{i2\pi(\mathbf{g}+n\mathbf{k})\cdot\mathbf{x}} = \mathrm{Q}_n\, \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} = \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} \tag{3.18}$$

for all $\mathbf{x} \in \mathbb{R}^d$ if $\mathbf{g} \in \mathbb{Z}^d_{n,\square}$ and $\mathbf{k} \in \mathbb{Z}^d$. Using these facts we get, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}
\mathrm{Q}_n\, u(\mathbf{x}) &= \mathrm{Q}_n \left(\sum_{\mathbf{m} \in \mathbb{Z}^d} [u]_{\mathbf{m}}\, \mathrm{e}^{i2\pi \mathbf{m}\cdot\mathbf{x}} \right) \\
&= \mathrm{Q}_n \left(\sum_{\mathbf{g} \in \mathbb{Z}^d_{n,\square}} \sum_{\mathbf{k} \in \mathbb{Z}^d} [u]_{\mathbf{g}+n\mathbf{k}}\, \mathrm{e}^{i2\pi (\mathbf{g}+n\mathbf{k})\cdot\mathbf{x}} \right) \\
&= \sum_{\mathbf{g} \in \mathbb{Z}^d_{n,\square}} \left(\sum_{\mathbf{k} \in \mathbb{Z}^d} [u]_{\mathbf{g}+n\mathbf{k}} \right) \mathrm{Q}_n\, \mathrm{e}^{i2\pi (\mathbf{g}+n\mathbf{k})\cdot\mathbf{x}} \\
&= \sum_{\mathbf{g} \in \mathbb{Z}^d_{n,\square}} \left(\sum_{\mathbf{k} \in \mathbb{Z}^d} [u]_{\mathbf{g}+n\mathbf{k}} \right) \mathrm{e}^{i2\pi \mathbf{g}\cdot\mathbf{x}} \qquad \text{by (3.18).}
\end{aligned}$$

Note that $\sum_{\mathbf{k} \in \mathbb{Z}^d} [u]_{\mathbf{g}+n\mathbf{k}}$ is well-defined for all $\mathbf{g} \in \mathbb{Z}^d_{n,\square}$ since the Fourier Series of u is convergent. □

We can now go on and present the following estimates for Q_n operating on continuous functions (recall from Theorem 3.27 that $H^t_p \subset C_p$ when $t > 1/2$ for $d = 1$ and when $t > 1$ for $d = 2$). These results can be found in [72].

Lemma 3.32. *The interpolation projection has the following approximation error bounds.*

1. *For $d = 1$, $t > 1/2$, $0 \le s \le t$ and $u \in H^t_p$ we have*

$$\|u - \mathrm{Q}_n\, u\|_{H^s_p} \le C_t \left(\tfrac{n}{2}\right)^{s-t} \|u\|_{H^t_p}$$

   where $C_t = (1 + \sum_{j=1}^{\infty} \frac{1}{j^{2t}})^{1/2}$.

2. *For $d = 2$, $t > 1$, $0 \le s \le t$ and $u \in H^t_p$ we have*

$$\|u - \mathrm{Q}_n\, u\|_{H^s_p} \le C_{s,t} \left(\tfrac{n}{2}\right)^{s-t} \|u\|_{H^t}$$

   where $C_{s,t} = (2^s \sum_{j,k=0}^{\infty} |j^2 + k^2|_\star^{-t})^{1/2}$.

Proof. Part 1 is Theorem 8.3.1 on page 243 of [72]. Part 2 is Theorem 8.5.3 on page 253 of [72]. □

3.3 Piecewise Continuous Functions

In this section we discuss definitions and regularity results for piecewise continuous functions. We also prove bounds on the Fourier coefficients of periodic piecewise continuous functions.

In the first subsection we define two spaces of periodic, piecewise continuous functions. For the rest of this thesis we restrict ourselves to these particular types of piecewise continuous functions. In the second subsection we prove regularity results for our periodic piecewise continuous functions and in the third subsection we bound the corresponding Fourier coefficients.

3.3.1 Two Special Classes of Periodic Piecewise Continuous Functions

In this section we use PC_p and PC_p' to denote spaces of periodic piecewise continuous functions.

For the case when $d = 1$, the definition of a piecewise continuous function on Ω is clear, although for Fourier Series results to hold we must restrict ourselves to functions with bounded variation.

When $d \geq 2$, we restrict ourselves to a special class of piecewise continuous functions such that the interfaces (sets where the funtion is discontinuous) can be described as the boundaries of Lipschitz domains.

For both cases, $d = 1$ and $d \geq 2$, we make a further restriction and specify that our piecewise continuous functions must also be bounded and infinitely differentiable on regions of continuity. This final restriction is not strictly necessary for Theorem 3.40. However, the proof is much easier since we can apply Lemma 3.38. A weaker condition for Theorem 3.40 would specify only finite differentiability in the regions of continuity where the order of differentiability depends on d.

We start by defining Lipschitz continuous, Liptshitz hypographs and Lipschitz domains (i.e. a domain with a Lipschitz boundary). We rely on the definitions on page 89 of [54].

Definition 3.33. For any domain $\Gamma \subseteq \mathbb{R}^d$, a function $f : \Gamma \to \mathbb{R}$ is called Lipschitz continuous if there exists a constant C such that

$$|f(x) - f(y)| \leq C|x - y| \qquad \forall x, y \in \Gamma.$$

Definition 3.34. Let $d \geq 2$ and let $\zeta : \mathbb{R}^{d-1} \to \mathbb{R}$ be a Lipschitz continuous function. Then the following set is a Liphshitz hypograph

$$\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}_d < \zeta(\mathbf{x}') \text{ for all } \mathbf{x}' = (\mathbf{x}_1, \ldots, \mathbf{x}_{d-1}) \in \mathbb{R}^{d-1}\}.$$

Definition 3.35. Let $d \geq 2$. The open set $\Gamma \subset \mathbb{R}^d$ is a Lipschitz domain if its boundary $\partial\Gamma$ is compact and if there exist finite families $\{V_j\}$ and $\{W_j\}$ that have the following properties:

1. The family $\{W_j\}$ is a finite open cover of $\partial\Gamma$, i.e., each W_j is an open subset of $\mathbb{R}^d$, and $\partial\Gamma \subseteq \bigcup_j W_j$.

2. Each $V_j \subset \mathbb{R}^d$ is a transformation by a rigid body motion of a Lipschitz hypograph, i.e. Each V_j can be transformed into a Lipschitz hypograph by rotation and translation. For later reference we will denote this transformation by $S : \mathbb{R}^d \to \mathbb{R}^d$ where S maps the Lipschitz hypograph to V_j.

3. V_j satisfies $W_j \cap \Gamma = W_j \cap V_j$ for each j.

See Figure 3-2 for an example of how W_j and V_j are defined.

For later reference, we make the remark here that $\partial\Gamma$ is a C^∞ class boundary if we replace Lipschitz hypographs with C^∞ hypographs ($\zeta \in C^\infty(\mathbb{R}^{d-1})$) in the definition above.

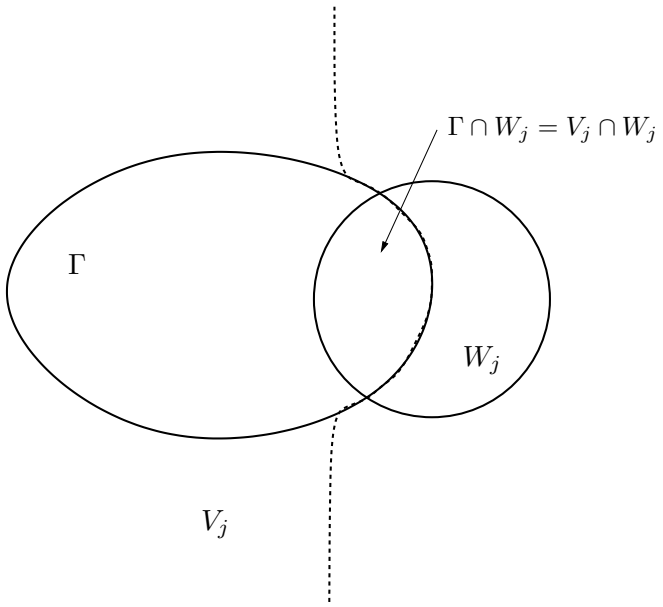

Figure 3-2: Diagram of a Lipschitz domain showing how the V_j and W_j are defined.

Now, using the definition of Lipschitz domains we define our special class of piecewise continuous functions using the following representation.

Definnition 3.36. For $d \in \mathbb{N}$ a periodic function f is in PC_p (our special class of periodic, piecewise continuous functions) if it can be represented in the following way:

$$f(\mathbf{x}) = f_0 + \sum_{j=1}^{J} f_j(\mathbf{x}) \qquad \forall \mathbf{x} \in \Omega \tag{3.19}$$

where $f_0 \in C_p^\infty \cap BV(\Omega)$ ($BV(\Omega)$ denotes the set of functions on Ω with bounded

variation) and $f_j(\mathbf{x})$ are periodic, piecewise continuous functions of the form

$$f_j(\mathbf{x}) = \begin{cases} c_j(\mathbf{x}) & \mathbf{x} \in \Omega_j \\ 0 & \mathbf{x} \in \overline{\Omega} \backslash \Omega_j \end{cases}$$

where each c_j is the restriction to Ω_j of a function in $C^\infty(\Omega) \cap BV(\Omega)$ and the Ω_j are a finite number of Lipschitz domains such that $\Omega_j \subset\subset \Omega$. The interfaces of $f(\mathbf{x})$ are the sets $\partial\Omega_j$.

Sometimes (in 2D) we will need to be more restrictive in our choice of periodic piecewise constant functions. In these cases we will use the following definition.

Definition 3.37. For $d = 2$, a periodic function f is in PC_p' if it is in PC_p with the additional assumption that each Ω_j is a convex Lipschitz polygon with a finite number of corners.

3.3.2 Regularity

In this section we prove the regularity of our special class of periodic, piecewise continuous functions. We begin by presenting two results from [54]. The first result proves the regularity of a simple discontinuous function where the discontinuity is on the boundary between two half spaces. This result is given as an exercise in [54] and we present the proof in the Appendix A.2. The second result, however, proves that we can distort our simple discontinuous function to a discontinuous function where the shape of the interface region can be represented with a Lipschitz continuous function and the regularity will be preserved. We do not prove the second result as it is proved in [54].

In the main theorem we will use a third result from [54] but we do not state it in a separate lemma.

Lemma 3.38. *Let $u \in C_0^\infty(\mathbb{R}^d)$ and define*

$$f(\mathbf{x}) := \begin{cases} u(\mathbf{x}) & x_d < 0 \\ 0 & x_d \geq 0 \end{cases}$$

Then $f \in H^{1/2-\epsilon}(\mathbb{R}^d)$ for any $\epsilon > 0$.

This result is based on exercise 3.22 on page 112 of [54]. We present the proof in Appendix A.2.

Now we quote Theorem 3.23 on page 85 of [54]. The proof is omitted as it is given in [54].

Lemma 3.39. *Suppose that $\kappa : \mathbb{R}^d \to \mathbb{R}^d$ is a bijective map and r is a positive integer such that $D^\alpha \kappa$ and $D^\alpha \kappa^{-1}$ exist and are (uniformly) Lipschitz on $\mathbb{R}^d$ for $|\alpha| \leq r - 1$.*

Then for $1 - r \leq s \leq r$ *we have*

$$u \in H^s(\mathbb{R}^d) \quad \iff \quad u \circ \kappa \in H^s(\mathbb{R}^d)$$

and in which case there exist constants $c, C > 0$ *(that depend on* κ*) such that*

$$c\|u\|_{H^s(\mathbb{R}^d)} \leq \|u \circ \kappa\|_{H^s(\mathbb{R}^d)} \leq C\|u\|_{H^s(\mathbb{R}^d)}$$

for all $u \in H^s(\mathbb{R}^d)$.

We now have the preliminary results from which we will develop our main theorem about the regularity of our special class of piecewise continuous functions.

Theorem 3.40. *Let* $f \in PC_p$ *(see Definition 3.36). Then for any* $\epsilon > 0$,

$$f \in H_p^{1/2-\epsilon}.$$

Proof. Let $s < 1/2$. Using the representation of f given in (3.19) we write

$$\|f\|_{H_p^s} \leq \|f_0\|_{H_p^s} + \sum_{j=1}^{J} \|f_j\|_{H_p^s}$$

Since $f_0 \in C_p^\infty$, $\|f_0\|_{H_p^s} < \infty$. We consider each $\|f_j\|_{H_p^s}$ separately. Recall that the Ω_j associated with f_j satisfy $\Omega_j \subset\subset \Omega$. Therefore, choose θ according to Lemma 3.17 so that $\theta(\mathbf{x}) = 1$ for $\mathbf{x} \in \Omega_j$. Also recall that Ω_j is a Lipschitz domain and according to the definition of a Lipschitz domain, there exists a finite open cover of $\partial\Omega_j$. Denote this by $\{W_k\}_{k=1}^K$. Define W_{K+1} to cover the interior of Ω_j such that $W_{K+1} \cap \partial\Omega_j = \emptyset$. The set $\{W_k\}_{k=1}^{K+1}$ is now a finite open cover of Ω_j. Now invoke Corollary 3.22 on page 84 of [54] to get a partition of unity, $\phi_1, \phi_2, \ldots, \phi_{K+1}$ for Ω_j such that $\phi_m \in C^\infty(\mathbb{R}^d)$ and $\operatorname{supp}\phi_m \subseteq W_m$ for every $m = 1, \ldots, K+1$, and $\sum_m \phi_m = 1$ on Ω_j. Using ϕ_m, θ and Lemma 3.29 we can write

$$\|f_j\|_{H_p^s} \leq C\|\theta f_j\|_{H^s(\mathbb{R}^d)} = \left\|\sum_{m=1}^{K+1} \phi_m \theta f_j\right\|_{H^s(\mathbb{R}^d)} \leq \sum_{m=1}^{K+1} \|\phi_m \theta f_j\|_{H^s(\mathbb{R}^d)}$$

Now treat each $\|\phi_m \theta f_j\|_{H^s(\mathbb{R}^d)}$ separately. We construct a bijective κ so that we can use Lemma 3.39. Define S to be the rotation and translation associated with W_m from Definition 3.35 and define $T : \mathbb{R}^d \to \mathbb{R}^d$ as a vertical shear, $T(\mathbf{x}) := (\mathbf{x}', x_d + \zeta(\mathbf{x}'))$ for all $\mathbf{x} \in \mathbb{R}^d$, where ζ is the Lipschitz continuous funciton used in the Lipschitz hypograph in Definition 3.35. Both S and T are bijective and Lipschitz so we can define $\kappa := S \circ T$ and κ is bijective and Lipschitz. Note that for $d = 1$ we define κ to shift the boundary

of Ω_j to the origin. Applying Lemma 3.39 with $r = 1$ we get

$$\|\phi_m \theta f_j\|_{H^s(\mathbb{R}^d)} \leq \tfrac{1}{c} \|(\phi_m \theta f_j) \circ \kappa\|_{H^s(\mathbb{R}^d)} \quad \text{for } 0 \leq s \leq 1.$$

Now we show that by the construction of κ, $(\phi_m \theta f_j) \circ \kappa$ satisfyies the assumptions of Lemma 3.38.

By the representation of f in (3.19) we see that f_j is a restriction to Ω_j of a function $g_j \in C^\infty(\mathbb{R}^d)$. Also, $\operatorname{supp} \phi_m \subseteq W_m$ implies that

$$\phi_m \theta f_j(\mathbf{x}) = \begin{cases} h_j(\mathbf{x}) & \mathbf{x} \in W_m \cap \Omega_j \\ 0 & \mathbf{x} \notin W_m \cap \Omega_j \end{cases}$$

where $h_j = \phi_m \theta g_j \in C_0^\infty(\mathbb{R}^d)$. Define $\kappa^{-1}(W_m) = \{\mathbf{y} \in \mathbb{R}^d : \kappa(\mathbf{y}) \in W_m\}$. By the definition of κ we have

$$\begin{aligned} \mathbf{x} \in W_m \cap \Omega_j \quad &\Longrightarrow \quad \mathbf{x} = \kappa(\mathbf{y}) \text{ for } \mathbf{y} \in \kappa^{-1}(W_m) \text{ with } y_d < 0 \\ \mathbf{x} \in W_m \cap (\mathbb{R}^d \backslash \Omega_j) \quad &\Longrightarrow \quad \mathbf{x} = \kappa(\mathbf{y}) \text{ for } \mathbf{y} \in \kappa^{-1}(W_m) \text{ with } y_d \geq 0 \end{aligned}$$

Therefore we have

$$(\phi_m \theta f_j) \circ \kappa(\mathbf{y}) = \begin{cases} h \circ \kappa(\mathbf{y}) & \mathbf{y} \in \{\kappa^{-1}(W_m) : y_d < 0\} \\ 0 & \mathbf{y} \in \{\kappa^{-1}(W_m) : y_d \geq 0\} \\ 0 & \mathbf{y} \notin \kappa^{-1}(W_m) \end{cases}$$

where $h \circ \kappa \in C_0^\infty(\mathbb{R}^d)$ and the assumptions of Lemma 3.38 are satisfied. Therefore, by Lemma 3.38,

$$\|\phi_m \theta f_j \circ \kappa\|_{H^{1/2-\epsilon}(\mathbb{R}^d)} < \infty$$

Since this statement holds for $m = 1, \ldots, M$, and $j = 1, \ldots, J$ our proof is complete. □

3.3.3 Fourier Coefficients

In this subsection we try to develop results that tell us about the behaviour of the Fourier coefficients of piecewise constant functions. We would like to estimate the Fourier coefficients of functions in our special class of periodic piecewise continuous functions, PC_p, that we defined in Definition 3.36. Unfortunately, for the case when $d = 2$ the best that we can do is estimate the Fourier coefficients of periodic piecewise continuous functions in PC_p'.

We begin with results for when $d = 1$ before considering the case when $d = 2$. The following result is a corollary of Theorem 39 on page 26 of [36] and can be proved using integration by parts.

Lemma 3.41. *If $f \in L_p^2$ is continuous on Ω except at a finite number of points where there is a jump and is absolutely continuous in the intervals of continuity then there exists a constant F such that*

$$|[f]_n| \leq F|n|^{-1} \qquad \forall n \in \mathbb{Z}, n \neq 0.$$

Proof. Suppose f has J discontinuities at $x_1, x_2, \ldots, x_J$ and let $d_j = f(x_j+0)-f(x_j-0)$ (i.e. let d_j be the size of the jump at each discontinuity). Assume for convenience and without loss of generality that $x_j \neq \pm\frac{1}{2}$. For $0 \neq n \in \mathbb{Z}$, subdividing Ω into intervals of continuity and integrating by parts yields

$$[f]_n = \frac{1}{i2\pi n}\sum_{j=1}^{J} d_j \,\mathrm{e}^{-i2\pi n x_j} + \frac{1}{i2\pi n}\int_\Omega f'(x)\,\mathrm{e}^{-i2\pi n x}\,dx$$

Since f is absolutely continuous on each interval of continuity, it is has bounded variation on each interval and therefore $f' \in L^1(\Omega)$ (see [4]). Therefore,

$$|[f]_n| \leq \frac{1}{2\pi n}\left(\sum_{j=1}^{J}|d_j| + \|f'\|_{L^1(\Omega)}\right)$$

□

Using this estimate for the coefficients of a piecewise continuous function (which requires slightly different assumptions on f) and the definition of H_p^s we can obtain an alternative proof for Theorem 3.40 (in the 1D case).

When $d = 2$ it is not so easy to estimate the asymptotic behaviour of the Fourier coefficients of a piecewise continuous function. Before we present our main theorem of this subsection let us present the following two illustrative examples.

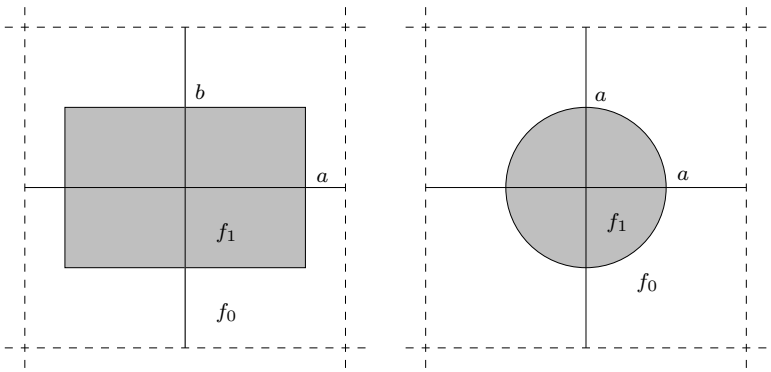

Figure 3-3: Diagram of $f(\mathbf{x})$ from Examples 3.42 (left) and 3.43 (right).

Example 3.42. Rectangular hole. For $0 < a, b < \frac{1}{2}$ and constants f_0 and f_1, define $f \in L_p^2$ by

$$f(\mathbf{x}) = \begin{cases} f_1 & |x_1| < a \text{ and } |x_2| < b \\ f_0 & \text{elsewhere in } \Omega. \end{cases}$$

See Figure 3-3. Then $f(\mathbf{x})$ has Fourier coefficients,

$$[f]_{\mathbf{g}} = \begin{cases} f_0 + (f_1 - f_0)ab & \mathbf{g} = \mathbf{0} \\ (f_1 - f_0)a\frac{\sin(g_2\pi b)}{g_2\pi} & g_1 = 0,\ g_2 \neq 0 \\ (f_1 - f_0)\frac{\sin(g_1\pi a)}{g_1\pi}b & g_1 \neq 0,\ g_2 = 0 \\ (f_1 - f_0)\frac{\sin(g_1\pi a)\sin(g_2\pi b)}{g_1 g_2 \pi^2} & g_1 \neq 0,\ g_2 \neq 0. \end{cases} \tag{3.20}$$

From (3.20) we can see that $|[f]_{\mathbf{g}}| \leq \frac{|f_1 - f_0|}{g_1 g_2}$ when $\mathbf{g}$ is not perpendicular to any of the interfaces of $f(\mathbf{x})$ and $|[f]_{\mathbf{g}}| \leq \frac{|f_1 - f_0|}{|g|}$ when $\mathbf{g}$ is perpendicular to the interfaces of $f(\mathbf{x})$. With these Fourier coefficients it is possible to prove that there exists a constant F such that

$$C_n = \left(\sum_{|g_1|+|g_2|=n} |[f]_{\mathbf{g}}|^2 \right)^{\frac{1}{2}} \leq Fn^{-1} \qquad n \in \mathbb{N}.$$

We do this using the following argument,

$$\begin{aligned} C_n^2 = \sum_{|g_1|+|g_2|=n} |[f]_{\mathbf{g}}|^2 &\leq (f_1 - f_0)^2 \left(\tfrac{4}{\pi^2 n^2} + \tfrac{4}{\pi^4} \sum_{k=1}^{n-1} \tfrac{1}{k^2(n-k)^2} \right) \\ &\leq \tfrac{4(f_1-f_0)^2}{\pi^2} \left(\tfrac{1}{n^2} + \tfrac{2}{\pi^2} \sum_{k=1}^{\lfloor n/2 \rfloor} \tfrac{1}{k^2(n/2)^2} \right) \\ &= \tfrac{4(f_1-f_0)^2}{\pi^2} \left(\tfrac{1}{n^2} + \tfrac{8}{\pi^2 n^2} \sum_{k=1}^{\lfloor n/2 \rfloor} \tfrac{1}{k^2} \right) \\ &\leq \tfrac{4(f_1-f_0)^2}{\pi^2} \left(\tfrac{1}{n^2} + \tfrac{8}{\pi^2 n^2} \tfrac{\pi^2}{6} \right) \\ &\leq \tfrac{28(f_1-f_0)^2}{3\pi^2} \tfrac{1}{n^2} \leq (f_1 - f_0)^2 n^{-2} \qquad n \in \mathbb{N}. \end{aligned}$$

For the next example, instead of having a rectangular interface, we work with a circular interface.

Example 3.43. Circular hole. For $0 < a < \frac{1}{2}$ and constants f_0 and f_1, define $f \in L_p^2$ by

$$f(\mathbf{x}) = \begin{cases} f_1 & |\mathbf{x}| < a \\ f_0 & \text{elsewhere in } \Omega. \end{cases}$$

See Figure 3-3. Then $f(\mathbf{x})$ has Fourier coefficients,

$$[f]_{\mathbf{g}} = \begin{cases} f_0 + (f_1 - f_0)\pi a^2 & \mathbf{g} = \mathbf{0} \\ (f_1 - f_0)\frac{a}{2\pi|\mathbf{g}|} J_1(\pi|\mathbf{g}|a) & |\mathbf{g}| \neq 0 \end{cases}$$

where J_1 is the 1st order Bessel function. With these Fourier coefficients we can again prove that there exists a constant F such that

$$C_n = \left(\sum_{|g_1|+|g_2|=n} |[f]_{\mathbf{g}}|^2 \right)^{\frac{1}{2}} \leq F n^{-1} \qquad n \in \mathbb{N}.$$

To prove this property we use the following argument. From the properties of Bessel functions, we know that there exists a constant A such that $J_1(r) \leq A r^{-1/2}$ for $r > 0$. Therefore, $|[f]_{\mathbf{g}}| \leq \frac{|f_1 - f_0|A}{2\pi^{3/2}} |\mathbf{g}|^{-3/2}$ and

$$\begin{aligned} C_n^2 &= \sum_{|g_1|+|g_2|=n} |[f]_{\mathbf{g}}|^2 \\ &\leq \frac{(f_1 - f_0)^2 A^2}{4\pi^3} \sum_{|g_1|+|g_2|=n} \frac{1}{|\mathbf{g}|^3} \\ &\leq \frac{(f_1 - f_0)^2 A^2}{4\pi^3} \frac{4n}{(n/\sqrt{2})^3} \\ &= \frac{4\sqrt{2}(f_1 - f_0)^2 A^2}{\pi^3} \frac{1}{n^2} = F^2 \frac{1}{n^2} \end{aligned}$$

Now let us state some Lemmas in preparation for the main theorem of this subsection.

Lemma 3.44. *Let $f \in H_p^s$ for $s \in \mathbb{R}$ and define $g \in H_p^s$ by $g(\mathbf{x}) = f(\mathbf{x} + \mathbf{x}_0)$ for $\mathbf{x}_0 \in \mathbb{R}^d$. Then*

$$[g]_{\mathbf{g}} = [f]_{\mathbf{g}} \, \mathrm{e}^{i2\pi \mathbf{g} \cdot \mathbf{x}_0} \qquad \forall \mathbf{g} \in \mathbb{Z}^d.$$

Proof.

$$g(\mathbf{x}) = f(\mathbf{x} + \mathbf{x}_0) = \sum_{\mathbf{g} \in \mathbb{Z}^d} [f]_{\mathbf{g}} \, \mathrm{e}^{i2\pi \mathbf{g} \cdot (\mathbf{x} + \mathbf{x}_0)} = \sum_{\mathbf{g} \in \mathbb{Z}^d} \left([f]_{\mathbf{g}} \, \mathrm{e}^{i2\pi \mathbf{g} \cdot \mathbf{x}_0} \right) \mathrm{e}^{i2\pi \mathbf{g} \cdot \mathbf{x}} \qquad \forall \mathbf{x} \in \mathbb{R}^d.$$

□

Before we state the next two lemmas let us recall that

$$|h|_\star = \begin{cases} |h| & \text{if } h \neq 0 \\ 1 & \text{if } h = 0 \end{cases}.$$

for all $h \in \mathbb{R}$.

Lemma 3.45. *Let $u \in C^\infty(\mathbb{R}^2)$ with $\operatorname{supp} u \subset \Omega$, let $\mathbf{x}_0 \in \Omega$ and $\mathbf{v} \in \mathbb{R}^2$, and define*

$$f(\mathbf{x}) := \begin{cases} u(\mathbf{x}) & (\mathbf{x}-\mathbf{x}_0)\cdot\mathbf{v} \le 0 \\ 0 & (\mathbf{x}-\mathbf{x}_0)\cdot\mathbf{v} > 0. \end{cases}$$

Also define $F(\mathbf{x}) = f(\mathbf{x}+\mathbf{x}_0)$ for all $\mathbf{x} \in \mathbb{R}^2$ and define $G(\mathbf{y}) = F(S(\mathbf{y}))$ for all $\mathbf{y} \in \mathbb{R}^2$ where S is a rotation such that $G(\mathbf{y}) = 0$ for all $y_2 > 0$. Then there exists a constant A such that

$$|[f]_{\mathbf{g}}| \le \frac{A}{|h_1|_\star |h_2|_\star} \qquad \forall \mathbf{g} \in \mathbb{Z}^2,\ \mathbf{h} = S^{-1}(\mathbf{g}).$$

Proof. Let $\mathbf{0} \neq \mathbf{g} \in \mathbb{Z}$. With the definitions in the lemma we get

$$\begin{aligned} |[f]_{\mathbf{g}}| &= \left| \int_\Omega f(\mathbf{x})\, \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \right| \\ &= \left| \int_{\mathbb{R}^2} f(\mathbf{x})\, \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \right| && \text{since } \operatorname{supp} f \subset \operatorname{supp} u \subset \Omega \\ &= \left| \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{x}_0} \int_{\mathbb{R}^2} F(\mathbf{x})\, \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \right| \\ &= \left| \int_{\mathbb{R}^2} F(\mathbf{x})\, \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \right| \\ &= \left| \int_{y_2<0} G(\mathbf{y})\, \mathrm{e}^{-i2\pi \mathbf{h}\cdot\mathbf{y}}\, d\mathbf{y} \right| && \text{with } \mathbf{x} = S(\mathbf{y}) \text{ and } \mathbf{h} = S^{-1}(\mathbf{g}). \end{aligned} \tag{3.21}$$

If $\mathbf{g}$ is *not* perpendicular or parallel to $\mathbf{v}$ then $h_1 \neq 0$ and $h_2 \neq 0$ and using (3.21) and integration by parts we get,

$$\begin{aligned} |[f]_{\mathbf{g}}| &= \left| \int_{y_2<0} G(\mathbf{y})\, \mathrm{e}^{-i2\pi \mathbf{h}\cdot\mathbf{y}}\, d\mathbf{y} \right| \\ &= \tfrac{1}{2\pi|h_1|} \left| \int_{y_2<0} (D_{y_1} G)(\mathbf{y})\, \mathrm{e}^{-i2\pi \mathbf{h}\cdot\mathbf{y}}\, d\mathbf{y} \right| \\ &= \tfrac{1}{4\pi^2|h_1||h_2|} \left| \int_{y_2=0} (D_{y_1} G)(\mathbf{y})\, \mathrm{e}^{-i2\pi \mathbf{h}\cdot\mathbf{y}}\, d\mathbf{y} \right| \\ &\qquad + \tfrac{1}{4\pi^2|h_1||h_2|} \left| \int_{y_2<0} (D_{y_2} D_{y_1} G)(\mathbf{y})\, \mathrm{e}^{-i2\pi \mathbf{h}\cdot\mathbf{y}}\, d\mathbf{y} \right| \\ &\le \tfrac{1}{4\pi^2|h_1||h_2|} \left(\int_{y_2=0} |D_{y_1} G_(\mathbf{y})| d\mathbf{y} + \int_{y_2<0} |D_{y_2} D_{y_1} G(\mathbf{y})| d\mathbf{y} \right) \\ &\le \frac{A}{|h_1||h_2|} \end{aligned} \tag{3.22}$$

Alternatively, if $\mathbf{g}$ is perpendicular to $\mathbf{v}$ ($h_2 = 0$) or if $\mathbf{g}$ is parallel to $\mathbf{v}$ ($h_1 = 0$) then we can not carry out both integrations by parts in (3.22) and we only get the following

estimate for $|[f]_{\mathbf{g}}|$ instead,

$$|[f]_{\mathbf{g}}| \leq \begin{cases} \frac{A}{|h_2|} & \text{if } h_1 = 0 \\ \frac{A}{|h_1|} & \text{if } h_2 = 0 \end{cases} \tag{3.23}$$

Hence, the result. Note that A is a generic constant that may be different for (3.22) and (3.23). □

Lemma 3.46. *Let $u \in C^\infty(\mathbb{R}^2)$ with $\operatorname{supp} u \subset \Omega$, let $\mathbf{x}_0 \in \Omega$ and $\mathbf{v}, \mathbf{w} \in \mathbb{R}^2$ such that $\mathbf{v} \neq k\mathbf{w}$ for all $k \in \mathbb{R}$. Define*

$$f(\mathbf{x}) := \begin{cases} u(\mathbf{x}) & \text{if } (\mathbf{x}-\mathbf{x}_0)\cdot\mathbf{v} \leq 0 \text{ and } (\mathbf{x}-\mathbf{x}_0)\cdot\mathbf{w} \leq 0 \\ 0 & \text{for all other } \mathbf{x}. \end{cases}$$

Also define $F(\mathbf{x}) = f(\mathbf{x}+\mathbf{x}_0)$ for all $\mathbf{x} \in \mathbb{R}^2$ and define $G(\mathbf{y}) = F(S(\mathbf{y}))$ for all $\mathbf{y} \in \mathbb{R}^2$ where S is a rotation such that the lines $S(\mathbf{v})\cdot\mathbf{y} = 0$ and $S(\mathbf{w})\cdot\mathbf{y} = 0$ correspond to the lines $y_1 = 0$ and $y_2 = cy_1$ for $-\infty < c < \infty$ and $G(\mathbf{y}) = 0$ for all $y_1 < 0$ and $y_2 > cy_1$. Then there exists a constant A such that

$$|[f]_{\mathbf{g}}| \leq \frac{A}{|h_1 + ch_2|_\star |h_2|_\star} \qquad \forall \mathbf{g} \in \mathbb{Z}^2,\ \mathbf{h} = S^{-1}(\mathbf{g}).$$

Proof. Let $\mathbf{0} \neq \mathbf{g} \in \mathbb{Z}$. With the definitions in the lemma and since $\operatorname{supp} f \subset \operatorname{supp} u \subset \Omega$ we get

$$\begin{aligned} |[f]_{\mathbf{g}}| &= \left| \int_\Omega f(\mathbf{x})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \right| = \left| \int_{\mathbb{R}^2} f(\mathbf{x})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \right| \\ &= \left| \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}_0} \int_{\mathbb{R}^2} F(\mathbf{x})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \right| = \left| \int_{\mathbb{R}^2} F(\mathbf{x})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \right| \\ &= \left| \int_{\substack{y_1>0 \\ y_2<cy_1}} G(\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{h}\cdot\mathbf{y}}\, d\mathbf{y} \right| \qquad \text{with } \mathbf{x} = S(\mathbf{y}) \text{ and } \mathbf{h} = S^{-1}(\mathbf{g}). \end{aligned} \tag{3.24}$$

If $\mathbf{g}$ is *not* parallel with $\mathbf{v}$ or $\mathbf{w}$ then $h_2 \neq 0$ and $h_1 + ch_2 \neq 0$ and using (3.24) and integration by parts we get,

$$\begin{aligned} |[f]_{\mathbf{g}}| &= \left| \int_{\substack{y_1>0 \\ y_2<cy_1}} G(\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{h}\cdot\mathbf{y}}\, d\mathbf{y} \right| \\ &\leq \tfrac{1}{2\pi|h_2|} \left(\left| \int_0^\infty \left[G(\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{h}\cdot\mathbf{y}} \right]_{y_2=\infty}^{y_2=cy_1} dy_1 \right| + \left| \int_{\substack{y_1>0 \\ y_2<cy_1}} D_{y_2} G(\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{h}\cdot\mathbf{y}}\, d\mathbf{y} \right| \right) \\ &= \tfrac{1}{2\pi|h_2|} \left(\left| \int_0^\infty G(y, cy)\, \mathrm{e}^{-i2\pi(h_1+ch_2)y}\, dy \right| + \left| \int_{\substack{y_1>0 \\ y_2<cy_1}} D_{y_2} G(\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{h}\cdot\mathbf{y}}\, d\mathbf{y} \right| \right). \end{aligned}$$

Continuing to integrate by parts we get,

$$\begin{aligned}|[f]_{\mathbf{g}}| &\leq \tfrac{1}{4\pi^2|h_2||h_1+ch_2|}\left(|G(\mathbf{0})| + \left|\int_0^\infty D(G(y,cy))\,\mathrm{e}^{-i2\pi(h_1+ch_2)y}\,dy\right|\right.\\ &\qquad \left. + \left|\int_{y_2=cy_1} D_{y_2}G(\mathbf{y})\,\mathrm{e}^{-i2\pi\mathbf{h}\cdot\mathbf{y}}\,d\mathbf{y}\right| + \left|\int_{\substack{y_1>0\\ y_2<cy_1}} D_{y_1}D_{y_2}G(\mathbf{y})\,\mathrm{e}^{-i2\pi\mathbf{h}\cdot\mathbf{y}}\,d\mathbf{y}\right|\right)\\ &\leq \frac{A}{|h_2||h_1+ch_2|}\end{aligned} \tag{3.25}$$

Note that A depends on f and c, and that c might be very large. In this sense (3.25) may not be a particularly sharp bound.

Alternatively, if $\mathbf{g}$ is parallel to $\mathbf{v}$ or $\mathbf{w}$ ($h_2 = 0$ or $h_1 + ch_2 = 0$) then we can not carry out both integrations by parts in (3.25) and we only get the following estimate for $|[f]_{\mathbf{g}}|$ instead,

$$|[f]_{\mathbf{g}}| \leq \begin{cases} \frac{A}{|h_2|} & \text{if } h_1 + ch_2 = 0 \\ \frac{A}{|h_1+ch_2|} & \text{if } h_2 = 0 \end{cases} \tag{3.26}$$

Hence, the result. Note that A is a generic constant that may be different for (3.25) and (3.26). □

Now we present the main theorem of this subsection (it is an original result). Unfortunately, our proof is limited to the function space PC'_p (see Definition 3.37) which is more restrictive than PC_p (see Definition 3.36), in that only convex Lipschitz polygon interfaces with a finite number of corners are permitted. However, we think it may be possible to extend our result to functions from PC_p that have Lipschitz interfaces.

Theorem 3.47. *Let $d = 2$ and assume that $f \in PC'_p$ (see Definition 3.37). Then there exists a constant F such that*

$$C_n = \left(\sum_{|g_1|+|g_2|=n} |[f]_{\mathbf{g}}|^2\right)^{\frac{1}{2}} \leq Fn^{-1} \qquad n \in \mathbb{N}.$$

Proof. Recall from Definition 3.37 (and Definition 3.36) that we can write

$$f(\mathbf{x}) = f_0(\mathbf{x}) + \sum_{j=1}^{J} f_j(\mathbf{x})$$

With $f(\mathbf{x})$ defined in this way we can split $|[f]_{\mathbf{g}}|$ into the following,

$$|[f]_{\mathbf{g}}| \leq |[f_0]_{\mathbf{g}}| + \sum_{j=1}^{J} |[f_j]_{\mathbf{g}}| \qquad \forall \mathbf{g} \in \mathbb{Z}^2. \tag{3.27}$$

Now let us fix j and consider each $|[f_j]_{\mathbf{g}}|$ separately. Since Ω_j (from definition of PC'_p) is a convex Lipschitz polygon with a finite number of corners we can define a finite open cover of $\partial\Omega_j$, $\{W_k\}_{k=1}^{K}$, such that each W_k covers either a single corner of $\partial\Omega_j$ or a straight edge of $\partial\Omega_j$. Also define W_{K+1} to cover the interior of Ω_j so that $W_{K+1} \cap \partial\Omega_j = \emptyset$. The family $\{W_k\}_{k=1}^{K+1}$ is a finite open cover of Ω_j. Now invoke Corollary 3.22 of [54] to get a partition of unity $\{\phi_k\}_{k=1}^{K+1}$ for Ω_j such that $\phi_k \in C^\infty(\mathbb{R}^2)$ and $\operatorname{supp}\phi_k \subset W_k$ for every $k = 1, \ldots, K+1$ and $\sum_k \phi_k = 1$ on Ω_j. Using our partition of unity and the definition of a Fourier coefficient (see Definition 3.13) we get

$$|[f_j]_{\mathbf{g}}| = \left| \sum_{k=1}^{K+1} [\phi_k f_j]_{\mathbf{g}} \right| \leq \sum_{k=1}^{K+1} |[\phi_k f_j]_{\mathbf{g}}| \qquad \mathbf{g} \in \mathbb{Z}^2. \tag{3.28}$$

With (3.27) and (3.28) we can write

$$\begin{aligned} C_n^2 &= \sum_{|g_1|+|g_2|=n} |[f]_{\mathbf{g}}|^2 \\ &\leq (J+1)^2 \sum_{|g_1|+|g_2|=n} \left(|[f_0]_{\mathbf{g}}|^2 + \sum_{j=1}^{J} |[f_j]_{\mathbf{g}}|^2 \right) \\ &\leq (J+1)^2 (K+1)^2 \sum_{|g_1|+|g_2|=n} \left(|[f_0]_{\mathbf{g}}|^2 + \sum_{j=1}^{J} \sum_{k=1}^{K+1} |[\phi_k f_j]_{\mathbf{g}}|^2 \right) \\ &\leq (J+1)^2 (K+1)^2 \Bigg(\underbrace{\sum_{|g_1|+|g_2|=n} |[f_0]_{\mathbf{g}}|^2}_{I_1(n)} + \sum_{j=1}^{J} \sum_{k=1}^{K+1} \underbrace{\sum_{|g_1|+|g_2|=n} |[\phi_k f_j]_{\mathbf{g}}|^2}_{I_2(n)} \Bigg). \end{aligned} \tag{3.29}$$

Now we bound $I_1(n)$ and $I_2(n)$ separately.

By Lemma 3.15 we know that there exists a constant A_0 such that $|[f_0]_{\mathbf{g}}| \leq A_0 |\mathbf{g}|^{-2}$ for every $\mathbf{0} \neq \mathbf{g} \in \mathbb{Z}^2$. Using this we bound $I_1(n)$ in the following way,

$$\begin{aligned} I_1(n) &= \sum_{|g_1|+|g_2|=n} |[f_0]_{\mathbf{g}}|^2 \leq A_0^2 \sum_{|g_1|+|g_2|=n} \frac{1}{|\mathbf{g}|^4} \leq A_0^2 (4n) \frac{1}{(n/\sqrt{2})^4} \\ &= \frac{16 A_0^2}{n^3} \leq \frac{16 A_0^2}{n^2} = B_0 n^{-2} \qquad \forall n \in \mathbb{N}. \end{aligned} \tag{3.30}$$

To bound $I_2(n)$ let us fix j and k and consider each $|[\phi_k f_j]_{\mathbf{g}}|$ separately. First, let us consider the case when $k = K+1$. In this case $\phi_{K+1} f_j \in C^\infty(\mathbb{R}^2)$ and $\operatorname{supp}(\phi_{K+1} f_j) \subset \Omega$ so we can extend $\phi_{K+1} f_j$ beyond Ω periodically and use Lemma 3.15 to show that there exists a constant $A_{j,K+1}$ such that $|[\phi_{K+1} f_j]_{\mathbf{g}}| \leq A_{j,K+1} |\mathbf{g}|^{-2}$ for every $\mathbf{0} \neq \mathbf{g} \in \mathbb{Z}^2$. Using the same argument as in (3.30) we can show that when $k = K+1$ there exists a constant $B_{j,K+1}$ such that $I_2(n) \leq B_{j,K+1} n^{-2}$.

Now let us consider $|[\phi_k f_j]_{\mathbf{g}}|$ for the cases when $k = 1, \ldots, K$. There are two

possibilities; either W_k covers a corner of $\partial\Omega_j$ or W_k covers a straight edge of $\partial\Omega_j$.

If W_k covers a straight edge of $\partial\Omega_j$, then $\phi_k f_j(\mathbf{x})$ has the form of $f(\mathbf{x})$ in Lemma 3.45. Therefore, applying Lemma 3.45, there exists a rotation $S = S_{jk}$ and a constant A_{jk} such that

$$|[\phi_k f_j]_{\mathbf{g}}| \leq \frac{A_{jk}}{|h_1|_\star |h_2|_\star} \qquad \forall \mathbf{g} \in \mathbb{Z}^2,\ \mathbf{h} = S^{-1}(\mathbf{g}). \tag{3.31}$$

Alternatively, if W_k covers a corner of $\partial\Omega_j$, then $\phi_k f_j(\mathbf{x})$ has the form of $f(\mathbf{x})$ in Lemma 3.46 (this is where we require that Ω_j is convex). Therefore, applying Lemma 3.46, there exists a rotation $S = S_{jk}$ and constants A_{jk} and $c = c_{jk}$ with $-\infty < c_{jk} < \infty$ such that

$$|[\phi_k f_j]_{\mathbf{g}}| \leq \frac{A_{jk}}{|h_1 + ch_2|_\star |h_2|_\star} \qquad \forall \mathbf{g} \in \mathbb{Z}^2,\ \mathbf{h} = S^{-1}(\mathbf{g}). \tag{3.32}$$

Now we will (3.32) to bound I_2 for the case when W_k covers a corner of $\partial\Omega_j$ (the straight edge case is a special case of the corner case with $c = 0$). In order to bound I_2 we will need to define the following four sets of points,

$$\begin{aligned}
\mathcal{U}_n &:= \{\mathbf{g} \in \mathbb{Z}^2 : |g_1| + |g_2| = n\} \\
\mathcal{V}_n &:= \{\mathbf{v} = k\mathbf{g} : |\mathbf{v}| = \tfrac{n}{\sqrt{2}}, \mathbf{g} \in \mathcal{U}_n, k \in \mathbb{R}, 0 < k \leq 1\} \\
\mathcal{W}_n &:= \{\mathbf{w} = S^{-1}(\mathbf{v}) : \mathbf{v} \in \mathcal{V}_n\} \\
\mathcal{X}_n &:= \{\mathbf{x} = k\mathbf{w} : |x_2| = d \text{ or } |x_1 + cx_2| = d\sqrt{1+c^2}, \mathbf{w} \in \mathcal{W}_n, k \in \mathbb{R}, 0 < k \leq 1\}
\end{aligned} \tag{3.33}$$

where $d = \frac{n}{\sqrt{2}} \min(1 + (\sqrt{1+c^2} \pm c)^2)^{-1/2}$. Note that the vectors in $\mathcal{U}_n$ lie on a rotated ($\frac{\pi}{4}$ radians) square with sides of length $\sqrt{2}n$ centred at the origin; the vectors in $\mathcal{V}_n$ lie on a circle with radius $\frac{n}{\sqrt{2}}$ centred at the origin; the vectors in $\mathcal{W}_n$ also lie on a circle with radius $\frac{n}{\sqrt{2}}$ centred at the origin; and d has been calculated so that the points in $\mathcal{X}_n$ lie on the largest possible rhombus inside a circle of radius $\frac{n}{\sqrt{2}}$ centred at the origin where the sides of the rhombus are perpendicular to either $(0, 1)$ or $(1, c)$. Also note that d is the closest distance that a point in $\mathcal{X}_n$ can be to the origin. Let us define α to be the smallest interior angle of the rhombus, then

$$\alpha = \begin{cases} \tan^{-1}(-\frac{1}{c}) & c \neq 0 \\ \frac{\pi}{2} & c = 0. \end{cases}$$

It is possible to define bijections between each of these sets. For example, each $\mathbf{v} \in \mathcal{V}_n$ is a scaled vector in $\mathcal{U}_n$, each $\mathbf{w} \in \mathcal{W}_n$ is a rotation of a vector in $\mathcal{V}_n$, and each $\mathbf{x} \in \mathcal{X}_n$ is a scaled vector in $\mathcal{W}_n$. All of the bijections preserve the relative angles between the vectors in each set. Moreover, we can bound (from above and below) the angle between neighbouring points using the following argument. If we consider the vectors in $\mathcal{U}_n$ then the smallest angle between neighbouring vectors will be equal to the angle

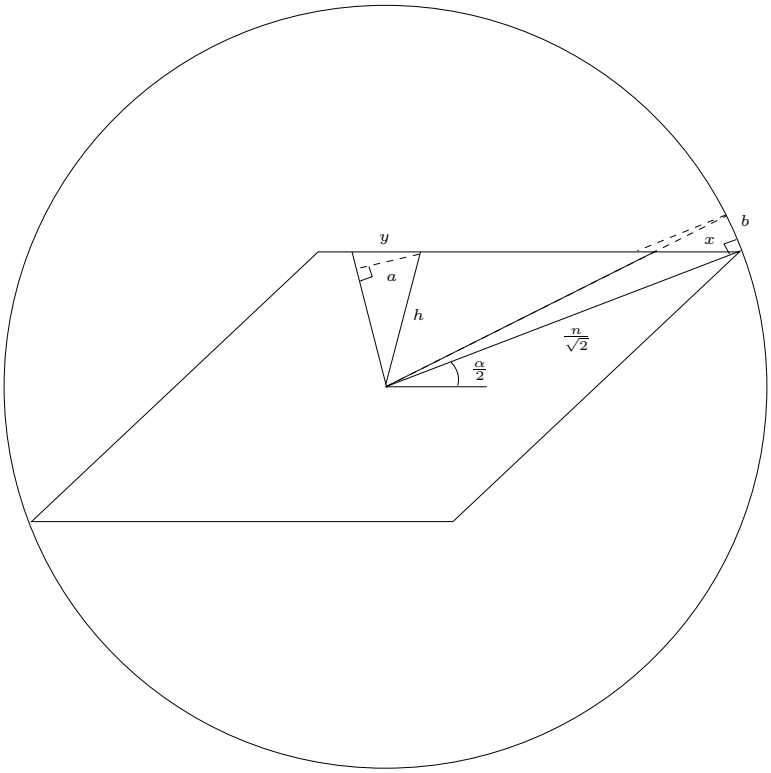

Figure 3-4: Diagram of a rhombus within a circle that correspond to the vectors in $\mathcal{W}_n$ and $\mathcal{X}_n$ respectively. x is the upper bound for the distance between two neighbouring vectors from $\mathcal{X}_n$ and y is the lower bound for the distance betwen two neighbouring vectors.

between $(0, n)$ and $(1, n-1)$, which is equal to $\tan^{-1}(\frac{1}{n-1})$. The largest angle between neighbouring vectors will be bounded above by two times the angle between $(\frac{n}{2}, \frac{n}{2})$ and $(\frac{n+1}{2}, \frac{n-1}{2})$, which is equal to $2\tan^{-1}(\frac{1}{n})$. Therefore, if θ is the angle between two neighbouring vectors in any of our sets then

$$\frac{1}{n-1} \leq \tan\theta \leq \frac{2n}{n^2-1}. \tag{3.34}$$

Note that in deriving (3.34) we used the identity $\tan(2A) = \frac{2\tan A}{1-\tan^2 A}$.

Now consider two neighbouring vectors in $\mathcal{X}_n$, $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ with $|x_2^{(1)}| = |x_2^{(2)}| = d$ and let θ denote the angle between them. We can use (3.34) to show that the distance between the two points is bounded above and below independently of n, i.e. we would like to bound $|x_1^{(1)} - x_1^{(2)}|$ from above and below.

As in Figure 3-4 we see that an upper bound will occur is y and the lower bound will be at x. Using simple trigonometry identities we get (for $n \geq 2$)

$$x \leq \frac{b}{\sin(\frac{\alpha}{2})} = \frac{\frac{n}{\sqrt{2}}\tan\theta}{\sin(\frac{\alpha}{2})} \leq \frac{n}{\sqrt{2}\sin(\frac{\alpha}{2})}\frac{2n}{n^2-1} \leq \frac{\sqrt{2}}{\sin(\frac{\alpha}{2})}(1 + \frac{1}{n^2-1}) \leq \frac{4\sqrt{2}}{3\sin(\frac{\alpha}{2})} =: A_x$$

and

$$y \geq a = h\sin\theta \geq d\sin\theta \geq d\sin(\tan^{-1}(\tfrac{1}{n-1})) = d\frac{\frac{1}{n-1}}{\sqrt{1+(\frac{1}{n-1})^2}} \geq d\tfrac{1}{n-1}$$
$$= \tfrac{1}{\sqrt{2}}\min(1+(\sqrt{1+c^2}\pm c)^2)^{-1/2}\tfrac{n}{n-1} \geq \tfrac{1}{\sqrt{2}}\min(1+(\sqrt{1+c^2}\pm c)^2)^{-1/2} := a_x.$$

Therefore, there exist positive constants a_x and A_x (independent of n) such that

$$a_x \leq |x_1^{(1)} - x_1^{(2)}| \leq A_x. \tag{3.35}$$

By symmetry we get an equivalent result for when the two neighbouring points satisfy $|x_1^{(1)} + cx_2^{(1)}| = |x_1^{(2)} + cx_2^{(2)}| = d\sqrt{1+c^2}$.

Now we can bound $I_2(n)$.

$$\begin{aligned}
I_2(n) &= \sum_{\mathbf{g}\in\mathcal{U}_n} |[\phi_k f_j]_{\mathbf{g}}|^2 \\
&\leq A_{jk}^2 \sum_{\mathbf{g}\in\mathcal{U}_n} \frac{1}{|h_1+ch_2|_\star^2|h_2|_\star^2} && \text{where } \mathbf{h}=S^{-1}(\mathbf{g}) \\
&\leq A_{jk}^2 \sum_{\mathbf{g}\in\mathcal{V}_n} \frac{1}{|h_1+ch_2|_\star^2|h_2|_\star^2} && \text{where } \mathbf{h}=S^{-1}(\mathbf{g}) \\
&= A_{jk}^2 \sum_{\mathbf{h}\in\mathcal{W}_n} \frac{1}{|h_1+ch_2|_\star^2|h_2|_\star^2} \\
&\leq A_{jk}^2 \sum_{\mathbf{h}\in\mathcal{X}_n} \frac{1}{|h_1+ch_2|_\star^2|h_2|_\star^2} \\
&= A_{jk}^2 \left(2\sum_{\substack{\mathbf{h}\in\mathcal{X}_n \\ |h_2|=d}} \frac{1}{|h_1+ch_2|_\star^2 d^2} + 2 \sum_{\substack{\mathbf{h}\in\mathcal{Y}_n \\ |h_1+ch_2|=d\sqrt{1+c^2}}} \frac{1}{d^2(1+c^2)|h_2|_\star^2} \right) \\
&\leq \frac{A_{jk}^2}{d^2}\left(8+8\sum_{k=1}^{\lceil n/2\rceil}\frac{1}{k^2 a_x^2|\sin\alpha|^2}\right) && \text{by (3.35) and symmetry} \\
&\leq \frac{8A_{jk}^2}{d^2}\left(1+\frac{1}{a_x^2|\sin\alpha|^2}\sum_{k=1}^{\infty}\frac{1}{k^2}\right) \\
&\leq \frac{16A_{jk}^2}{n^2\min(1+(\sqrt{1+c^2}\pm c)^2)^{-1/2}}\left(1+\frac{\pi^2}{6a_x^2|\sin\alpha|^2}\right) \\
&= B_{jk}n^{-2} && \text{for } n\geq 2.
\end{aligned}$$

To recap, we have shown that there exists constants B_0 and B_{jk} such that $I_1(n) \leq B_0 n^{-2}$ and $I_2(n) \leq B_{jk}n^{-2}$ (for both the straight edge and corner cases) for all $n \in \mathbb{N}$, $n \geq 2$. Therefore, putting these results back into (3.29) yields the result. □

We can use Lemma 3.47 to obtain an alternative proof for Theorem 3.40 in the case when $f \in PC'_p$.

Lemma 3.48. *Let $d = 2$. For $f \in L^2_p$, if there exists a constant F such that the Fourier coefficients of f satisfy*

$$C_n = \left(\sum_{|g_1|+|g_2|=n} |[f]_{\mathbf{g}}|^2 \right)^{1/2} \leq Fn^{-1} \qquad n \in \mathbb{N},$$

Then

$$f \in H^{1/2-\epsilon}_p$$

for every $\epsilon > 0$.

Proof. Assume that there exists a constant F such that $C_n \leq Fn^{-1}$ for all $n \in \mathbb{N}$. Then, for any $\epsilon > 0$,

$$\begin{aligned} \|f\|^2_{H^{1/2-\epsilon}_p} &= \sum_{\mathbf{g}\in\mathbb{Z}^2} |\mathbf{g}|^{1-2\epsilon}_\star |[f]_{\mathbf{g}}|^2 \\ &\leq |[f]_0|^2 + \sum_{n=1}^{\infty} n^{1-2\epsilon} C_n^2 \\ &\leq |[f]_0|^2 + F \sum_{n=1}^{\infty} n^{-1-2\epsilon} < \infty. \end{aligned}$$

□

3.4 Operator and Spectral Theory

In this section we present the operator and spectral theory that we will need for studying photonic crystal fibres in the later chapters of this thesis. We will only be considering linear operators in this thesis. In the first subsection we will define compact, symmetric and self-adjoint operators on Hilbert spaces. We also define an extension of an operator and the adjoint of an operator. In another subsection we will then define the spectrum of an operator on a Hilbert space and we will present definitions that will let us characterise the spectrum of a self-adjoint operator as either *essential spectrum* or *discrete spectrum*. Next, we will present several results that will tell us when the spectrum of an operator is real and/or discrete. For completeness, we also define *absolutely continuous spectrum* since we have already used this term earlier in this thesis. Unfortunately, the definition of absolutely continuous spectrum is quite complicated and we will need to invoke the spectral theorem.

Following the subsection on the spectra of operators we will present a subsection on the Floquet Transform. We will present a definition of the Floquet Transform as

well as the key spectral result that relates the spectrum of an operator to the union of spectra of a family of operators obtained under the Floquet Transform.

The main reference for the spectral theory is [37] but we also use [42], whereas Floquet Theory references include [17], [44], [45] and [69].

Before we proceed, let us define a Hilbert space.

Definition 3.49. A Hilbert space $\mathcal{H}$ is a linear vector space with an inner product $(\cdot,\cdot)_{\mathcal{H}}$ that satisfies, for any $u, v, w \in \mathcal{H}$ and $a \in \mathbb{C}$,

1. $(au + v, w)_{\mathcal{H}} = a(u, w)_{\mathcal{H}} + (v, w)_{\mathcal{H}}$;
2. $(u, v)_{\mathcal{H}} = \overline{(v, u)_{\mathcal{H}}}$;
3. $(u, u)_{\mathcal{H}} \geq 0$ and $(u, u)_{\mathcal{H}} = 0 \Leftrightarrow u = 0$.

A Hilbert space $\mathcal{H}$ is also complete with respect to the norm induced by its inner product, $\|\cdot\|_{\mathcal{H}} = (\cdot,\cdot)_{\mathcal{H}}^{1/2}$.

All Hilbert spaces are also reflexive Banach spaces.

3.4.1 Operator Definitions

In this subsection we will present definitions for linear operators on Hilbert Spaces. We define the adjoint and the extension of an operator as well as what constitutes a symmetric or self-adjoint operator. We then define a compact operator a on Hilbert space.

Let A be a linear operator from a Hilbert space $\mathcal{H}_1$ to another Hilbert Space $\mathcal{H}_2$. By this we mean that A has domain $D(A) \subset \mathcal{H}_1$ that is dense in $\mathcal{H}_1$. The following defines the adjoint of A.

Definition 3.50. The adjoint of A, A^*, is a linear operator from $\mathcal{H}_2$ to $\mathcal{H}_1$ with domain

$$D(A^*) := \{v \in \mathcal{H}_2 : \exists v^* \in \mathcal{H}_1 \text{ such that } (Au, v)_{\mathcal{H}_2} = (u, v^*)_{\mathcal{H}_1} \forall u \in D(A)\}.$$

$D(A)$ dense in $\mathcal{H}_1$ implies that for every $v \in D(A^*)$ there exists a unique v^* such that $(Au, v)_{\mathcal{H}_2} = (u, v^*)_{\mathcal{H}_1}$ for all $u \in D(A)$ and we define $A^*v = v^*$. In particular,

$$(Au, v)_{\mathcal{H}_2} = (u, A^*v)_{\mathcal{H}_1} \qquad \forall u \in D(A), v \in D(A^*).$$

Now we define an extension of a linear operator on a Hilbert space.

Definition 3.51. Let $A_1 : \mathcal{H}_1 \to \mathcal{H}_2$ and $A_2 : \mathcal{H}_1 \to \mathcal{H}_2$ be two linear operators. A_2 is an extension of A_1 ($A_1 \subset A_2$) if

$$D(A_1) \subset D(A_2) \text{ and } A_1 u = A_2 u \quad \forall u \in D(A_1)$$

Using our definition of an extension of an operator it is easy to define a symmetric operator.

Definition 3.52. A linear operator $A : \mathcal{H} \to \mathcal{H}$ is symmetric if

$$(Au, v)_{\mathcal{H}} = (u, Av)_{\mathcal{H}} \qquad \forall u, v \in D(A).$$

This is equivalent to saying $A \subset A^*$.

Finally, we define a self-adjoint linear operator. Note that the condition A is self-adjoint is stronger than the condition A is symmetric.

Definition 3.53. A linear operator $A : \mathcal{H} \to \mathcal{H}$ is self-adjoint if $A = A^*$.

We see from the definition of a symmetric operator that A is self-adjoint if A is symmetric *and* $D(A) = D(A^*)$. This is the criterion we will use to show that an operator is self-adjoint in later chapters. We also remark that if A is bounded and symmetric then A is self-adjoint. We now define compact linear operators.

Definition 3.54. A bounded linear operator, A, on a reflexive Banach space is called compact if it maps a weakly convergent sequence into a strongly convergent sequence.

3.4.2 Spectra

In this subsection we define the spectrum of a linear operator on a Hilbert space. We will then define how to split the spectrum into two parts, the essential spectrum and the discrete spectrum. We then present some results that will tell when the spectrum of an operator is real and/or discrete. Finally, we present the definition of absolutely continuous spectrum.

We define the spectrum of an operator by first defining the resolvent set.

Definition 3.55. Let $A : \mathcal{H} \to \mathcal{H}$ be a linear (possibly unbounded) operator with domain $D(A)$ and let $\lambda \in \mathbb{C}$. λ is in the resolvent set, $\rho(A)$, if the operator $R_A(\lambda) := (A - \lambda)^{-1}$

1. exists;
2. the domain of $R_A(\lambda)$ is dense in $\mathcal{H}$; and
3. $R_A(\lambda)$ is bounded.

$R_A(\lambda)$ is called the resolvent of A. The spectrum of A, $\sigma(A)$, is defined by

$$\sigma(A) := \mathbb{C} \backslash \rho(A)$$

According to this definition we can classify $\lambda \in \sigma(A)$ depending on how the resolvent fails to satisfy the three conditions in the definition above. If $\lambda \in \sigma(A)$ then either,

1. $\ker(A-\lambda) \neq \{0\}$, or
2. $\ker(A-\lambda) = \{0\}$ but the range of $(A-\lambda)$ is not dense in $\mathcal{H}$ (in this case $R_A(\lambda)$ exists on the range of $(A-\lambda)$ but can not be *uniquely* extended to a bounded operator on $\mathcal{H}$), or
3. $\ker(A-\lambda) = \{0\}$ and the range of $(A-\lambda)$ is dense in $\mathcal{H}$ but $R_A(\lambda)$ is unbounded.

Definition 3.56. Let $A : \mathcal{H} \to \mathcal{H}$ be a linear operator and let $\lambda \in \sigma(A)$. In case 1. above ($\ker(A-\lambda) \neq \{0\}$) λ is called an *eigenvalue*, $u \in \ker(A-\lambda)$ is called an *eigenvector* or *eigenfunction* and $Au = \lambda u$. Moreover:

1. There is a smallest integer α, called the *ascent* of $A-\lambda$ such that $\ker(A-\lambda)^\alpha = \ker(A-\lambda)^{\alpha+1}$.
2. The functions in $\ker(A-\lambda)^\alpha$ are called *generalized eigenfunctions of A corresponding to* λ and the *order of a generalized eigenfunction* u is the smallest integer j such that $u \in \ker(A-\lambda)^j$.
3. The *geometric multiplicty of* A is equal to $\dim(\ker(A-\lambda))$.
4. The *algebraic multiplicity of* A is equal to $\dim(\ker(A-\lambda)^\alpha)$.

Note that the algebraic multiplicity is always greater than or equal to the geometric multiplicity.

Although we have made the distinction between algebraic and geometric multiplicity in the preceding definition, we will mostly work with compact, self-adjoint operators on Hilbert spaces. In this case, the ascent is 1, the algebraic multiplicity is equal to the geometric multiplicity and all generalised eigenfunctions are eigenfunctions in the usual sense, see page 683 in [6].

Now we split the spectrum into the discrete spectrum and the essential spectrum.

Definition 3.57. Let $A : \mathcal{H} \to \mathcal{H}$ be a linear operator. The discrete spectrum of A, $\sigma_d(A)$, is the set of all eigenvalues with finite (algebraic) multiplicity that are isolated points of $\sigma(A)$. The essential spectrum of A, $\sigma_{ess}(A)$, is defined by $\sigma_{ess}(A) = \sigma(A) \backslash \sigma_d(A)$.

It follows from this definition that if an eigenvalue with finite multiplicity is not isolated then it is in the essential spectrum. For self-adjoint operators we can chacterise the essential spectrum in terms of a Weyl sequence. See Definition 7.1 and Theorem 7.2 in [37].

Definition 3.58. Let $A : \mathcal{H} \to \mathcal{H}$ be a linear self-adjoint operator. A sequence $\{u_n\}$ is called a *Weyl sequence* for A and λ if $u_n \in D(A)$, $\|u_n\|_{\mathcal{H}} = 1$, $(u_n, v)_{\mathcal{H}} \to 0$ for all $v \in \mathcal{H}$ and $\|(A - \lambda)u_n\|_{\mathcal{H}} \to 0$ as $n \to \infty$.

Theorem 3.59. *Let $A : \mathcal{H} \to \mathcal{H}$ be a linear self-adjoint operator. Then $\lambda \in \sigma_{ess}(A)$ if and only if there exists a Weyl sequence for A and λ.*

Now we present four results that will be useful later in this thesis.

Theorem 3.60. *Let $A : \mathcal{H} \to \mathcal{H}$ be linear self-adjoint operator. Then:*

1. *$\sigma(A) \subset \mathbb{R}$.*
2. *If A is compact then $\sigma(A)$ consists of nonzero isolated eigenvalues of finite multiplicity with the only possible accumulation point at zero, and possibly zero (which may have infinite multiplicity).*
3. *If there exists a $\mu \in \rho(A)$ such that $R_A(\mu)$ is a compact operator, then $\sigma(A) = \sigma_d(A)$.*
4. *If $C : \mathcal{H} \to \mathcal{H}$ is compact operator, then $\sigma_{ess}(A) = \sigma_{ess}(A + C)$.*

Proof. The first result is a standard spectral theory result and is given in Theorem 5.5 on page 51 of [37].

Part 2 is Theorem 9.10 on page 93 of [37]. It is often called the Riesz-Schauder Theorem.

Part 3 follows from Part 2. By definition $R_A(\mu)$ is bounded, and since $R_A(\mu)$ is compact, Part 2 implies that the spectrum of $R_A(\mu)$ is a sequence eigenvalues $\lambda_1, \lambda_2, \ldots$ with finite multiplicity such that $|\lambda_n| \to 0$. $R_A(\mu)$ has a well-defined inverse, $A - \mu$, and so 0 is not an eigenvalue of $R_A(\mu)$. For $\lambda \in \sigma(R_A(\mu))$, let u be a corresponding eigenfunction. Then

$$\begin{aligned}(A - \mu)^{-1}u &= \lambda u \\ u &= \lambda(A - \mu)u \\ Au &= \left(\mu + \tfrac{1}{\lambda}\right) u.\end{aligned}$$

Therefore, $\mu + \frac{1}{\lambda}$ is an eigenvalue of A with corresponding eigenfunction u. Therefore, the spectrum of A consists of only isolated eigenvalues and the spectrum of A is discrete.

Part 4 is the classical Weyl Theorem as given on page 117 of [69]. □

The eigenfunctions of a compact, self-adjoint operator can also be characterised in a special way. The following theorem is Theorem 2.36 on page 47 of [54].

Theorem 3.61. *If $A : \mathcal{H} \to \mathcal{H}$ is compact and self-adjoint then there exist (possibly finite) sequences of functions $u_1, u_2, \dots$ in $\mathcal{H}$ and real numbers $\lambda_1, \lambda_2, \dots$ that have the following properties:*

1. *Each u_j is an eigenfunction of A with eigenvalue λ_j.*
2. *The eigenfunctions $u_1, u_2, \dots$ are orthonormal.*
3. *The eigenvalues satisfy $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$.*
4. *If the sequences are infinite then $\lambda_j \to 0$ as $j \to \infty$.*
5. *The set $U = \text{span}\{u_1, u_2, \dots\}$ is dense in $\mathcal{H}$.*

Now we use the spectral theorem (for example, see Chapter VI, Section 5.3 of [42]) to define the absolutely continuous spectrum of a linear self-adjoint operator. For a linear self-adjoint operator $A : \mathcal{H} \to \mathcal{H}$ the spectral theorem says that we can uniquely represent A by

$$A = \int_{-\infty}^{\infty} \lambda \, dE(\lambda) \tag{3.36}$$

where $\{E(\lambda) : -\infty < \lambda < \infty\}$ is a family of self-adjoint projection operators on $\mathcal{H}$ that satisfy

1. $E(\lambda)E(\mu) = E(\mu)E(\lambda) = E(\min\{\lambda, \mu\})$ for all $\lambda, \mu \in \mathbb{R}$,
2. $E(\lambda) = E(\lambda + 0)$ for all $\lambda \in \mathbb{R}$, i.e $E(\lambda)f = \lim_{\epsilon \searrow 0} E(\lambda + \epsilon)f$ for all $f \in \mathcal{H}$.
3. $\lim_{\lambda \to -\infty} E(\lambda) = 0$ and $\lim_{\lambda \to +\infty} E(\lambda) = I$.
4. If $S = (\lambda_1, \lambda_2] \subset \mathbb{R}$, with $E(S) := E(\lambda_2) - E(\lambda_1)$, $\lambda_1(f, f)_{\mathcal{H}} \leq (Af, f)_{\mathcal{H}} \leq \lambda_2(f, f)_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $\|(A - \lambda)f\|_{\mathcal{H}} \leq |\lambda_1 - \lambda_2| \|f\|_{\mathcal{H}}$ for all $\lambda \in S$ and $f \in \mathcal{H}$.
5. $f \in D(A) \Leftrightarrow \int_{-\infty}^{\infty} \lambda^2 d(E(\lambda)f, f)_{\mathcal{H}} = \int_{-\infty}^{\infty} \lambda^2 d\|E(\lambda)f\|_{\mathcal{H}} < \infty$, and if $f \in D(A)$ then $Af = \int_{-\infty}^{\infty} \lambda \, d(E(\lambda)f)$.

In fact, $E(S)$ is defined for all Borel sets S of the real line and for any $u \in \mathcal{H}$, $m_u(S) := (E(S)u, u)_{\mathcal{H}} = \|E(S)u\|_{\mathcal{H}}^2$ is a non-negative countably additive measure defined for Borel sets S (see page 516 of [42]). If $m_u(S)$ is absolutely continuous (with respect to Lebesgue measure $|S|$) we say that u *is absolutely continuous with respect to* A. The set of all $u \in \mathcal{H}$ which are absolutely continuous with respect to A is denoted $\mathcal{H}_{ac}$. Theorems 1.5 and 1.6 on pages 516 and 517 of [42] imply that we can consider the *part* of A corresponding to $\mathcal{H}_{ac}$, denoted by A_{ac}, i.e we define $D(A_{ac}) = \{f \in \mathcal{H}_{ac} \cap D(A) : Af \in \mathcal{H}_{ac}\}$ and $A_{ac}f := Af$ for all $f \in D(A_{ac})$. The absolutely continuous spectrum of A is then defined as $\sigma_{ac}(A) := \sigma(A_{ac})$.

As we have seen, the definition of absolutely continuous spectrum is quite technical. All we need to know, however, is that $\sigma_{ac}(A) \subset \sigma_{ess}(A)$ (see [42] page 519). Therefore, if an operator has purely absolutely continuous spectrum then it must have purely essential spectrum.

3.4.3 Floquet Transform

In this subsection we define the Floquet Transform. There are two versions that are used in the literature and they are very closely related. We will define only one version in this thesis as this is all we will need. We will also present the main theorem from Floquet Theory.

The Floquet Transform is used to transform an operator with periodic coefficients (period cell is $\Omega = [-\frac{1}{2}, \frac{1}{2}]^d$) operating on $L^2(\mathbb{R}^d)$ into a family of operators operating on L^2_p.

One definition of the Floquet Transform is the following.

Definition 3.62. Let $v \in L^2(\mathbb{R}^d)$ with $d = 1, 2$. The *Floquet Transform* of $v(\mathbf{x})$ at $\mathbf{x} \in \mathbb{R}^d$ is defined as

$$\mathcal{F}v(\mathbf{x}, \boldsymbol{\xi}) = \sum_{\mathbf{r} \in \mathbb{Z}^d} v(\mathbf{x} - \mathbf{r})\, \mathrm{e}^{-i\boldsymbol{\xi}\cdot(\mathbf{x}-\mathbf{r})} \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d$$

For any fixed $\boldsymbol{\xi} \in \mathbb{R}^d$, $\mathcal{F}v(\cdot, \boldsymbol{\xi})$ is a periodic function and $\mathcal{F}v(\cdot, \boldsymbol{\xi}) \in L^2_p$, whereas for any fixed $\mathbf{x} \in \mathbb{R}^d$, $\mathcal{F}v(\mathbf{x}, \cdot)$ is quasi-periodic, i.e.

$$\mathcal{F}v(\mathbf{x}, \boldsymbol{\xi} + 2\pi\mathbf{e}_j) = \mathrm{e}^{-i2\pi\mathbf{x}\cdot\mathbf{e}_j}\, \mathcal{F}v(\mathbf{x}, \boldsymbol{\xi}) \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d, \forall j \in \{1, \ldots, d\}$$

with period cell $B := [-\pi, \pi]^d$, where $\mathbf{e}_j$ is a unit vector in the j^{th} coordinate direction. Therefore, it is sufficient to know $\mathcal{F}v(\mathbf{x}, \boldsymbol{\xi})$ for $\mathbf{x} \in \Omega$ and $\boldsymbol{\xi} \in B$.

We now consider the action of the Floquet transform on periodic differential operators L with the form,

$$L := L(\mathbf{x}, \nabla) := \sum_{i,j=1}^{n} D_{x_j} a^{ij}(\mathbf{x}) D_{x_i} + \sum_{i=1}^{n} b^i(\mathbf{x}) D_{x_i} + c(\mathbf{x})$$

where the coefficients $a^{ij}(\mathbf{x}), b^i(\mathbf{x})$ and $c(\mathbf{x})$ are all periodic functions with period cell Ω. It is easy to show that for $v \in L^2(\mathbb{R}^d)$ and fixed $\mathbf{x} \in \mathbb{R}^d$ we get

$$\mathcal{F}(Lv)(\mathbf{x}, \boldsymbol{\xi}) = L(\mathbf{x}, \nabla_{\mathbf{x}} + \boldsymbol{\xi})\mathcal{F}v(\mathbf{x}, \boldsymbol{\xi}) \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d.$$

where

$$L(\mathbf{x}, \nabla_{\mathbf{x}} + \boldsymbol{\xi}) := \sum_{i,j=1}^{n} (D_{x_j} + \xi_j) a^{ij}(\mathbf{x})(D_{x_i} + \xi_i) + \sum_{i=1}^{n} b^i(\mathbf{x})(D_{x_i} + \xi_i) + c(\mathbf{x}).$$

We use the notation $L_{\boldsymbol{\xi}} = L(\mathbf{x}, \nabla + \boldsymbol{\xi})$.

Instead of considering the Floquet Transform for fixed $\mathbf{x} \in \Omega$, we will consider the transform for fixed $\boldsymbol{\xi} \in B$. In this way, we say that the operator L (with periodic coefficients) operating on $L^2(\mathbb{R}^d)$ is transformed into a family of operators $L_{\boldsymbol{\xi}}$ for $\boldsymbol{\xi} \in B$ where each operator $L_{\boldsymbol{\xi}}$ operates on L^2_p.

We relate the spectrum of the original operator with the spectra of our family of operators by stating the key result from Floquet Theory. The result and references to the proof can be found in [45].

Theorem 3.63. *If L is self-adjoint with periodic coefficients then*

$$\sigma(L) = \bigcup_{\boldsymbol{\xi} \in B} \sigma(L_{\boldsymbol{\xi}})$$

The proof follows from the notion that the Floquet transform expands L operating on $L^2(\mathbb{R}^d)$ into the *direct integral* (see page 281 of [69]) of operators $L_{\boldsymbol{\xi}}$ on the torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$.

3.5 Some Results from Functional Analysis

In this section we present some results from functional analysis for studying linear differential operators in the weak sense. Our aim is to estimate the eigenvalue and eigenfunction errors for approximating the solution to a variational eigenvalue problem.

In the first subsection we begin by considering the bounded linear operator T and a family of bounded linear operators T_n such that $\mathrm{T}_n \to \mathrm{T}$ in norm as $n \to \infty$. We condense the theory in [6] to write down error bounds for the eigenvalues and eigenfunctions of the operator T_n in terms of the difference between T and T_n.

In the second subsection we define a variational eigenvalue problem and the corresponding *solution operator*. We relate the spectrum of the variational eigenvalue problem to the spectrum of the solution operator.

In the third subsection we apply the Galerkin method to a variational eigenvalue problem and we construct a family of solution operators T_n that approximate the solution operator of the original variational eigenvalue problem. We bound the difference between these operators in terms of the approximation error.

In the fourth subsection we present Strang's First Lemma in a general setting. Strang's First Lemma is a result that we use for estimating errors introduced by small

modifications of the original problem (e.g. through smoothing discontinuous coefficients).

Finally, in the fifth subsection we consider a second order PDE boundary value problem. We write it in variational form and then we prove regularity results for the solution when we have smooth coefficients. The regularity results for the variational eigenvalue problems we will be solving will depend on these regularity theorems for boundary value problems. Understanding the regularity of the eigenfunctions of our variational eigenvalue problems will be the key to making sharp estimates.

Before we begin the first subsection let us make the following definitions. Throughout this section let $\mathcal{H}$ denote a Hilbert space with inner product $(\cdot,\cdot)$ and induced norm $\|\cdot\|$.

Eigenfunction errors will be measured in terms of the difference between eigenspaces. To measure the difference between eigenspaces that are subspaces of $\mathcal{H}$ we will rely on the following definition.

Definition 3.64. Let X and Y be two closed subspaces of $\mathcal{H}$. The *gap between X and Y* is defined as

$$\delta(X,Y) = \sup_{x\in X,\|x\|=1} \operatorname{dist}(x,Y) = \sup_{y\in Y,\|y\|=1} \operatorname{dist}(y,Y)$$

Later in this thesis, when we use $\delta(\cdot,\cdot)$ $\mathcal{H}$ will be the Hilbert space H^1_p.

Lemma 3.65. *Let X, Y and Z be closed subspaces of $\mathcal{H}$. Then*

$$\delta(X,Z) \le \delta(X,Y) + \delta(Y,Z)$$

The proof of this result is given in Appendix A.3.

From the second subsection onwards we will need the following definitions of properties for bilinear forms.

Definition 3.66. A bilinear form $a(\cdot,\cdot) : \mathcal{H}\times\mathcal{H} \to \mathbb{R}$ may be:

1. *bounded* if there exists a constant $C_b > 0$ such that

$$|a(u,v)| \le C_b\|u\|\|v\| \qquad \forall u,v\in\mathcal{H}; \tag{3.37}$$

2. *coercive* if there exists a constant $C_c > 0$ such that

$$a(v,v) \ge C_c\|v\|^2 \qquad \forall v\in\mathcal{H};\ \text{and} \tag{3.38}$$

3. *Hermitian* if

$$a(u,v) = \overline{a(v,u)} \qquad \forall u,v\in\mathcal{H}.$$

If a bilinear form satisfies all of the properties of Definition 3.66 then we get the following Lemma.

Lemma 3.67. *If a bilinear form $a(\cdot,\cdot)$ is bounded, coercive and Hermitian on $\mathcal{H}$ then it defines an inner product on $\mathcal{H}$ and its induced norm $|a(\cdot,\cdot)|^{1/2}$ is equivalent to $\|\cdot\|$.*

3.5.1 Error Bounds for Operators

In this subsection we consider a family of bounded linear operators T_n ($n \in \mathbb{N}$) such that $\mathrm{T}_n \to \mathrm{T}$ in norm as $n \to \infty$. We present a result that first establishes that the eigenvalues and eigenfunctions of T_n approximate those of T and then estimate the errors for these approximate eigenvalues and eigenfunctions in terms of the difference between the operators T and T_n. The result is a condensed version of the theory in [6].

Based on [6, Theorem 7.1 on page 685] and [6, Theorem 7.3 on page 689] we get the following result.

Theorem 3.68. *Let the following conditions hold:*

1. *$\mathrm{T} : \mathcal{H} \to \mathcal{H}$ is a bounded, linear, compact operator.*
2. *$\mathrm{T}_n : \mathcal{H} \to \mathcal{H}$ is a family of bounded, linear, compact operators such that $\|\mathrm{T} - \mathrm{T}_n\| \to 0$ as $n \to \infty$.*
3. *μ is an eigenvalue of T with (algebraic) multiplicity m, and corresponding eigenspace $M := \ker(\mu - \mathrm{T})^\alpha$ where α denotes the ascent of $(\mu - \mathrm{T})$.*

Then, for sufficiently large n, there exist m eigenvalues of T_n (counted according to algebraic multiplicities), $\mu_1(n), \ldots, \mu_m(n)$ with corresponding generalised eigenspaces $\mathcal{M}_1(n), \ldots, \mathcal{M}_m(n)$ and a space

$$\mathcal{M} = \bigoplus_{j=1}^{m} \mathcal{M}_j$$

such that

$$\delta(M, \mathcal{M}) \lesssim \|(\mathrm{T} - \mathrm{T}_n)|_M\|$$

and

$$|\mu - \mu_j| \lesssim \left\{ \sum_{i,k=1}^{m} |((\mathrm{T} - \mathrm{T}_n)\phi_i, \phi_k^*)| + \|(\mathrm{T} - \mathrm{T}_n)|_M\| \|(\mathrm{T}^* - \mathrm{T}_n^*)|_M\| \right\}^{\frac{1}{\alpha}}$$

for $j = 1, \ldots, m$, where $\{\phi_1, \ldots, \phi_m\}$ is a basis for M, T^* and T_n^* are the adjoints of T and T_n respectively, and $\{\phi_1^*, \ldots, \phi_m^*\}$ is a basis for the generalised eigenspace of T^*.

Note that in the theorem above, the eigenspaces $\mathcal{M}_1, \ldots, \mathcal{M}_m$ are spaces that include generalised eigenfunctions, i.e. $\mathcal{M}_j := \ker(\mu_j - \mathrm{T}_n)^{\alpha_j}$ where α_j is the ascent of μ_j for $j = 1, \ldots, m$. Throughout this thesis we will only be working with the case when the ascent of μ is one. This will usually be because T is compact and self-adjoint on a Hilbert space and so M will not contain any generalised eigenfunctions. When the ascent of μ is one, the algebraic multiplicity of μ is equal to the geometric multiplicity of μ.

When T or T_n are self-adjoint then Theorem 3.68 can be written down in a more simple form.

3.5.2 Variational Eigenvalue Problems

In this subsection we define a variational eigenvalue problem and we define the solution operator that corresponds to the bilinear forms from the variational eigenvalue problem. We then show the relationship between the solution operator and the variational eigenvalue problem.

Definition 3.69. A variational eigenvalue problem on $\mathcal{H}$ is defined as: Find an eigenvalue $\lambda \in \mathbb{C}$ and a non-zero eigenfunction $u \in \mathcal{H}$ such that

$$a(u, v) = \lambda b(u, v) \qquad \forall v \in \mathcal{H} \tag{3.39}$$

where $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are bilinear forms on $\mathcal{H}$.

Associated with the bilinear forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ in Definition 3.69 we define an operator that we call the *solution operator*.

Definition 3.70. Assume that $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are bounded bilinear forms, $a(\cdot, \cdot)$ is coercive and let $f \in \mathcal{H}$. Then $\mathrm{T}\, f$ is uniquely defined by

$$a(\mathrm{T}\, f, v) = b(f, v) \quad \forall v \in \mathcal{H} \tag{3.40}$$

In this way we have defined an operator $\mathrm{T} : \mathcal{H} \to \mathcal{H}$. We call T the *solution operator corresponding to* $a(\cdot, \cdot)$ *and* $b(\cdot, \cdot)$.

Sometimes, we will refer to T as the solution operator corresponding to a variational eigenvalue problem. We really mean that T is the solution operator corresponding to the bilinear forms in the variational eigenvalue problem.

The operator T is well-defined and bounded due to the Lax-Milgram Lemma. When $a(\cdot, \cdot)$ is Hermitian then T is self-adjoint. The compactness of T depends on properties of the Hilbert space $\mathcal{H}$.

The following lemma gives us the link between eigenpairs of the variational eigenvalue problem and eigenpairs of its associated solution operator.

Lemma 3.71. *(λ, u) is an eigenpair of the variational eigenvalue problem (3.39) with $\lambda \neq 0$, if and only if $(\frac{1}{\lambda}, u)$ is an eigenpair of the solution operator T corresponding to (3.39).*

Proof. Let (λ, u) be an eigenpair of (3.39) with $\lambda \neq 0$. Then

$$\begin{aligned} a(u,v) = \lambda b(u,v)\ \forall v \in \mathcal{H} \quad &\Leftrightarrow \quad a\left(\tfrac{1}{\lambda}u, v\right) = b(u,v)\ \forall v \in \mathcal{H} && \text{divide through by } \lambda \\ &\Leftrightarrow \qquad \mathrm{T}\,u = \tfrac{1}{\lambda}u && \text{by Definition 3.70.} \end{aligned}$$

□

Since the eigenpairs of the variational eigenvalue problem and the solution operator are linked, the idea of ascent, generalised eigenfunctions, algebraic multiplicty and geometric multiplicity for the variational eigenvalue problem are inherited from the solution operator.

3.5.3 Galerkin Method and Error Estimates

In this subsection we apply the Galerkin method to the variational eigenvalue problem (3.39) to get a discrete variational eigenvalue problem. We then define a solution operator that corresponds to the discrete variational eigenvalue problem before we bound the difference between the solution operator corresponding to the original problem and the new solution operator corresponding to the discrete problem in terms of the approximation error using Cea's Lemma.

Error estimates for the Galerkin method applied to (3.39) in terms of the approximation error then follow from Theorem 3.68 and Lemma 3.71.

We now define the Galerkin method.

Definition 3.72. For $n \in \mathbb{N}$ choose a finite dimensional subspace $\mathcal{V}_n \subset \mathcal{H}$. The *Galerkin method* applied to the variational eigenvalue problem (3.39) is: Find $\lambda_n \in \mathbb{C}$ and non-zero $u_n \in \mathcal{V}_n$ such that

$$a(u_n, v) = \lambda_n b(u_n, v) \qquad \forall v \in \mathcal{V}_n. \tag{3.41}$$

We call this problem the discrete variational eigenvalue problem.

The Galerkin method is defined by the choice of finite dimensional space $\mathcal{V}_n$. Associated with the choice of $\mathcal{V}_n$ is the approximation error. We define it as follows.

Definition 3.73. Let $u \in \mathcal{H}$. The *approximation error* of $\mathcal{V}_n$ associated with u is defined as

$$\inf_{\chi \in \mathcal{V}_n} \|u - \chi\|. \tag{3.42}$$

We want to choose a sequence of $\mathcal{V}_n$ so that the approximation error will tend to zero as $n \to \infty$.

Just as we defined a solution operator corresponding to (3.39) we can define a family of solution operators corresponding to (3.41) for $n \in \mathbb{N}$. Assuming that $a(\cdot,\cdot)$ is bounded and coercive and that $b(\cdot,\cdot)$ is bounded then for each $n \in \mathbb{N}$ and $f \in \mathcal{H}$ we can uniquely define $\mathrm{T}_n f \in \mathcal{V}_n$ by

$$a(\mathrm{T}_n f, v) = b(f, v) \quad \forall v \in \mathcal{V}_n. \tag{3.43}$$

In this way, for each $n \in \mathbb{N}$, we have defined an operator $\mathrm{T}_n : \mathcal{H} \to \mathcal{V}_n$.

We now prove several properties of T_n in the following Lemma. Notice that Parts 2 and 3 are estimates for the right-hand-sides in Theorem 3.68 in terms of the approximation error. Part 2 is Cea's Lemma.

Lemma 3.74. *Assume that $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are both bounded bilinear forms and that $a(\cdot,\cdot)$ is coercive according to Definition 3.66. Let T and T_n denote the solution operators associated with* (3.39) *and* (3.41) *respectively. Then the following properties hold:*

1. $\mathrm{T}_n = \mathrm{P}_n \mathrm{T}$ *where P_n is the projection from $\mathcal{H}$ onto $\mathcal{V}_n$ defined by*

$$a(\mathrm{P}_n u - u, v) = 0 \qquad \forall u \in \mathcal{H}, \forall v \in \mathcal{V}_n.$$

2. *For any $u \in \mathcal{H}$,*

$$\| \mathrm{T}_n u - \mathrm{T} u\| \leq \left(1 + \tfrac{C_b}{C_c}\right) \inf_{\chi \in \mathcal{V}_n} \| \mathrm{T} u - \chi\|$$

3. *For any $u, v \in \mathcal{H}$,*

$$|a(\mathrm{T} u - \mathrm{T}_n u, v)| \leq C_b \left(1 + \tfrac{C_b}{C_c}\right) \inf_{\chi \in \mathcal{V}_n} \| \mathrm{T} u - \chi\| \inf_{\chi \in \mathcal{V}_n} \|v - \chi\|$$

where C_b and C_c are the constants from (3.37) *and* (3.38) *associated with the bilinear form $a(\cdot,\cdot)$.*

Proof. Part 1. For any $u \in \mathcal{H}$ and any $v_n \in \mathcal{V}_n$ we get

$$\begin{aligned}
a(\mathrm{T}_n u, v_n) &= b(u, v_n) && \text{by definition of } \mathrm{T}_n \\
&= a(\mathrm{T} u, v_n) && \text{by definition of } \mathrm{T} \\
&= a(\mathrm{P}_n \mathrm{T} u, v_n) + a(\mathrm{T} u - \mathrm{P}_n \mathrm{T} u, v) \\
&= a(\mathrm{P}_n \mathrm{T} u, v_n) && \text{by definition of } \mathrm{P}_n.
\end{aligned}$$

It then follows that

$$\begin{aligned} & a((\mathrm{T}_n - \mathrm{P}_n\,\mathrm{T})u, v_n) = 0 && \forall u \in \mathcal{H}, \forall v_n \in \mathcal{V}_n \\ \implies \quad & a((\mathrm{T}_n - \mathrm{P}_n\,\mathrm{T})u, (\mathrm{T}_n - \mathrm{P}_n\,\mathrm{T})u) = 0 && \text{choosing } v_n = (\mathrm{T}_n - \mathrm{P}_n\,\mathrm{T})u \\ \implies \quad & \|(\mathrm{T}_n - \mathrm{P}_n\,\mathrm{T})u\| = 0 && \text{by coercivity of } a(\cdot,\cdot). \end{aligned}$$

The final statement is true for all $u \in \mathcal{H}$ and so $\mathrm{T}_n = \mathrm{P}_n\,\mathrm{T}$.
Part 2. This is just Cea's Lemma. Let $u \in \mathcal{H}$. Using $\mathcal{V}_n \subset \mathcal{H}$, subtract (3.43) from (3.40) to get

$$a(\mathrm{T}\,u - \mathrm{T}_n\,u, v_n) = 0 \qquad \forall v_n \in \mathcal{V}_n. \tag{3.44}$$

For all $v_n, w_n \in \mathcal{V}_n$ and using (3.44) we then get

$$\begin{aligned} a(\mathrm{T}\,u - w_n, v_n) &= a(\mathrm{T}\,u - \mathrm{T}_n\,u, v_n) + a(\mathrm{T}_n\,u - w_n, v_n) \\ &= a(\mathrm{T}_n\,u - w_n, v_n). \end{aligned} \tag{3.45}$$

Now choose $w_n \in \mathcal{V}_n$ such that $\mathrm{T}_n\,u - w_n \neq 0$. We get

$$\begin{aligned} \|\,\mathrm{T}_n\,u - w_n\| &\leq \tfrac{1}{C_c}\frac{|a(\mathrm{T}_n\,u - w_n, \mathrm{T}_n\,u - w_n)|}{\|\,\mathrm{T}_n\,u - w_n\|} && \text{by coercivity of } a(\cdot,\cdot) \\ &\leq \tfrac{1}{C_c}\sup_{0\neq v_n\in\mathcal{V}_n}\frac{|a(\mathrm{T}_n\,u - w_n, v_n)|}{\|v_n\|} \\ &= \tfrac{1}{C_c}\sup_{0\neq v_n\in\mathcal{V}_n}\frac{|a(\mathrm{T}\,u - w_n, v_n)|}{\|v_n\|} && \text{by (3.45)} \\ &\leq \tfrac{C_b}{C_c}\|\,\mathrm{T}\,u - w_n\| && \text{by boundedness of } a(\cdot,\cdot) \end{aligned}$$

Notice that this statement still holds if $\mathrm{T}_n\,u - w_n = 0$. Therefore,

$$\begin{aligned} \|\,\mathrm{T}\,u - \mathrm{T}_n\,u\| &\leq \|\,\mathrm{T}\,u - w_n\| + \|\,\mathrm{T}_n\,u - w_n\| \\ &\leq \left(1 + \tfrac{C_b}{C_c}\right)\|\,\mathrm{T}\,u - w_n\| \qquad \forall w_n \in \mathcal{V}_n \end{aligned}$$

The result follows by taking the infimum over $w_n \in \mathcal{V}_n$.

Part 3. This result is an adaptation of part of a proof in [6]. With $u, v \in \mathcal{H}$ we get

$$\begin{aligned} |a((\mathrm{T} - \mathrm{T}_n)u, v)| &= |a((\mathrm{T} - \mathrm{T}_n)u, v - \chi)| \qquad \forall \chi \in \mathcal{V}_n \qquad \text{by (3.44)} \\ &\leq C_b\|(\mathrm{T} - \mathrm{T}_n)u\|\|v - \chi\| \quad \text{by boundedness of } a(\cdot,\cdot) \\ &= C_b\|(\mathrm{T} - \mathrm{T}_n)u\| \inf_{\chi\in\mathcal{V}_n}\|v - \chi\| \\ &\leq C_b\left(1 + \tfrac{C_b}{C_c}\right)\inf_{\chi\in\mathcal{V}_n}\|\,\mathrm{T}\,u - \chi\| \inf_{\chi\in\mathcal{V}_n}\|v - \chi\| \qquad \text{by Part 2.} \end{aligned}$$

□

In later chapters of this thesis we will use these three properties in conjuction with Theorem 3.68 to develop the error analysis of the spectral Galerkin method.

3.5.4 Strang's First Lemma

In this subsection we present Strang's First Lemma (as in Theorem 4.1.1 on page 186 of [9]). In [9], Strang's (first) Lemma is used to obtain error estimates when numerical integration is needed to evaluate the bilinear form $a(\cdot,\cdot)$ to determine the entries of the coefficient matrix. By using a quadrature formula to evaluate $a(\cdot,\cdot)$ we effectively solve a discrete variational problem with a different bilinear form $\tilde{a}(\cdot,\cdot)$. By solving a modified problem we have introduced an additional error and Strang's Lemma bounds this error in terms of the difference between $a(\cdot,\cdot)$ and $\tilde{a}(\cdot,\cdot)$.

Here, we will not be using quadrature to evaluate $a(\cdot,\cdot)$. However, we will be using a modified bilinear form $\tilde{a}(\cdot,\cdot)$ instead of $a(\cdot,\cdot)$ when we apply the smoothing method, where the discontinuous coefficients of our problem are replaced with smooth coefficients. We are interested in bounding the error that we introduce by using this modified bilinear form.

It is important to note that in the following theorem $\mathcal{V} \subset \mathcal{H}$ is not necessarily a finite dimensional subspace. Indeed, we will apply the result when $\mathcal{V}$ is infinite dimensional.

Theorem 3.75. *Let $u \in \mathcal{H}$ be the solution to*

$$a(u,v) = F(v) \qquad \forall v \in \mathcal{H}$$

where $a(\cdot,\cdot)$ is a bounded, coercive bilinear form and $F(\cdot)$ is a bounded linear functional on $\mathcal{H}$. Also let $\mathcal{V} \subset \mathcal{H}$ and let $\tilde{u} \in \mathcal{V}$ be the solution to

$$\tilde{a}(\tilde{u},v) = \tilde{F}(v) \qquad \forall v \in \mathcal{V}$$

where $\tilde{a}(\cdot,\cdot)$ is a bilinear form that is coercive on $\mathcal{V}$ and $\tilde{F}(\cdot)$ is a bounded linear functional on $\mathcal{V}$. Then

$$\|u-\tilde{u}\| \leq C\left(\inf_{v\in\mathcal{V}}\left\{\|u-v\| + \sup_{w\in\mathcal{V}}\frac{|a(v,w)-\tilde{a}(v,w)|}{\|w\|}\right\} + \sup_{w\in\mathcal{V}}\frac{|F(w)-\tilde{F}(w)|}{\|w\|}\right)$$

where $C = \max(\frac{1}{\tilde{C}_c}, 1+\frac{C_b}{\tilde{C}_c})$, C_b is the constant in (3.37) corresponding to $a(\cdot,\cdot)$ and $\tilde{C}_c$ corresponds to $\tilde{a}(\cdot,\cdot)$ in (3.38).

Proof. Let $v \in \mathcal{V}$ such that $v \neq \tilde{u}$. Then we may write

$$\|u-\tilde{u}\| \leq \|u-v\| + \|v-\tilde{u}\| \tag{3.46}$$

Now set $0 \neq w = \tilde{u} - v \in \mathcal{V}$. Then, using the coercivity of $\tilde{a}(\cdot,\cdot)$ and the boundedness

of $a(\cdot,\cdot)$, we get

$$\begin{aligned}\tilde{C}_c\|\tilde{u}-v\|^2 &\leq \tilde{a}(\tilde{u}-v,\tilde{u}-v)\\ &= \tilde{a}(\tilde{u}-v,w)\\ &= a(u-v,w)+[a(v,w)-\tilde{a}(v,w)]+[\tilde{a}(\tilde{u},w)-a(u,w)]\\ &= a(u-v,w)+[a(v,w)-\tilde{a}(v,w)]+[\tilde{F}(w)-F(w)]\\ &\leq C_b\|u-v\|\|w\|+[a(v,w)-\tilde{a}(v,w)]+[\tilde{F}(w)-F(w)]\end{aligned}$$

Now divide through by $\tilde{C}_c\|\tilde{u}-v\| = \tilde{C}_c\|w\|$ to get

$$\|\tilde{u}-v\| \leq \tfrac{C_b}{\tilde{C}_c}\|u-v\| + \tfrac{1}{\tilde{C}_c}\frac{|a(v,w)-\tilde{a}(v,w)|}{\|w\|} + \tfrac{1}{\tilde{C}_c}\frac{|\tilde{F}(w)-F(w)|}{\|w\|}. \tag{3.47}$$

Now take the supremum over $w \in \mathcal{V}$ to get

$$\|\tilde{u}-v\| \leq \tfrac{C_b}{\tilde{C}_c}\|u-v\| + \tfrac{1}{\tilde{C}_c}\sup_{w\in\mathcal{V}}\frac{|a(v,w)-\tilde{a}(v,w)|}{\|w\|} + \tfrac{1}{\tilde{C}_c}\sup_{w\in\mathcal{V}}\frac{|\tilde{F}(w)-F(w)|}{\|w\|} \tag{3.48}$$

Notice that if (3.48) also holds if $v = \tilde{u}$ ($w = 0$). Now put (3.46) and (3.48) together and take the infimum over $v \in \mathcal{V}$ to get

$$\begin{aligned}\|u-\tilde{u}\| \leq \inf_{v\in\mathcal{V}}\left\{\left(1+\tfrac{C_b}{\tilde{C}_c}\right)\|u-v\| + \tfrac{1}{\tilde{C}_c}\sup_{w\in\mathcal{V}}\frac{|a(v,w)-\tilde{a}(v,w)|}{\|w\|}\right\}\\ + \tfrac{1}{\tilde{C}_c}\sup_{w\in\mathcal{V}}\frac{|F(w)-\tilde{F}(w)|}{\|w\|}\end{aligned}$$

□

3.5.5 Regularity

In this subsection we consider two second order elliptic PDE boundary value problems and we develop a regularity result with estimates for each problem. The first problem we look at will be posed on a bounded domain with homogeneous Dirichlet boundary conditions while the second problem will have periodic boundary conditons and periodic coefficients. Both problems will have smooth coefficients as well as other restrictions that we will assume.

We will use the regularity result for the periodic boundary value problem to obtain the regularity of periodic eigenfunctions for the same differential operators in later chapters.

Let $\Omega' \subset \mathbb{R}^d$ be a bounded open set such that $\partial\Omega'$ is of class C^∞ (see remark after

Definition 3.35). Consider an elliptic boundary value problem for $f \in L^2(\Omega')$,

$$\begin{aligned} Lu &= f \qquad \text{in } \Omega' \\ u &= 0 \qquad \text{on } \partial\Omega' \end{aligned} \tag{3.49}$$

where

$$L := -\sum_{i,j=1}^{d} D_{x_j}(a^{ij}(\mathbf{x})D_{x_i}) + \sum_{i=1}^{d} b^i(\mathbf{x})D_{x_i} + c(\mathbf{x}), \tag{3.50}$$

with coefficients $a^{ij}, b^i, c \in C^\infty(\Omega')$ that satisfy

$$\sum_{i,j=1}^{d} a^{ij}(\mathbf{x})\boldsymbol{\xi}_i\boldsymbol{\xi}_j \geq C|\boldsymbol{\xi}|^2 \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d, \mathbf{x} \in \Omega'$$

for some constant $C > 0$ (this is the definition of elliptic). We also restrict the coefficients so that L is self-adjoint in the sense defined in [52]. This requires that $a^{ij} = \overline{a^{ji}}$, $b^i = -\overline{b^i}$ and $c = \overline{c} - \sum_{i=1}^{d} \overline{D_{x_i} b^i}$ for all $i, j = 1, \ldots, d$. (The adjoint problem also has homogeneous Dirichlet boundary conditions).

In the usual way, we solve (3.49) in the weak sense. This leads to the variational problem: find $u \in H_0^1(\Omega')$ such that

$$a(u, v) = F(v) \qquad \forall v \in H_0^1(\Omega') \tag{3.51}$$

where $a(\cdot, \cdot)$ is a bilinear form and $F(v)$ is a linear functional, given by

$$\begin{aligned} a(u, v) &:= \int_{\Omega'} \sum_{i,j=1}^{d} a^{ij} D_{x_i} u \overline{D_{x_j} v} + \sum_{i=1}^{d} b^i D_{x_i} u \overline{v} + cu\overline{v} d\mathbf{x} \\ F(v) &:= (f, v)_{L^2(\Omega')}. \end{aligned}$$

The assumptions on the coefficients of L that we have given above imply that $a(\cdot, \cdot)$ is a bounded and coercive bilinear form and we can condense the result given on pages 188 and 189 of [52] to get the following result.

Theorem 3.76. *Consider the problem* (3.49) *with all of the restrictions on the coefficients listed above. Let $s \in \mathbb{R}$, with $s \geq 2$ and let $f \in H^{s-2}(\Omega')$. Then there exists a unique solution u to* (3.51) *such that $u \in H^s(\Omega')$ and*

$$\|u\|_{H^s(\Omega')} \leq C\|f\|_{H^{s-2}(\Omega')}$$

for a constant C (independent of f).

Proof. The uniqueness of the solution follows from the Lax-Milgram Lemma whereas the regularity and estimate come from the result on pages 188 and 189 of [52]. □

Now let us consider a boundary value problem with periodic boundary conditions. With period cell Ω defined as in previous sections, the boundary value problem with periodic boundary conditions for $f \in L^2_p$ is

$$\begin{aligned} Lu &= f \qquad \text{in } \mathbb{R}^d \\ u &\text{ is periodic with period cell } \Omega \end{aligned} \tag{3.52}$$

where L has the same expression as in (3.50) except we now assume that $a^{ij}, b^i, c \in C^\infty_p$, and

$$\sum_{i,j=1}^{d} a^{ij}(\mathbf{x})\boldsymbol{\xi}_i\boldsymbol{\xi}_j \geq C|\boldsymbol{\xi}|^2 \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d, \mathbf{x} \in \mathbb{R}^d$$

for some constant $C > 0$. We still require that $a^{ij} = \overline{a^{ji}}$, $b^i = -\overline{b^i}$ and $c = \overline{c} - \sum_{i=1}^d \overline{D_{x_i} b^i}$ for all $i, j = 1, \ldots, d$. The weak form of this problem is to search for $u \in H^1_p$ such that

$$a(u, v) = F(v) \qquad \forall v \in H^1_p. \tag{3.53}$$

where

$$\begin{aligned} a(u,v) &:= \int_\Omega \sum_{i,j=1}^{d} a^{ij} D_{x_i} u \overline{D_{x_j} v} + \sum_{i=1}^{d} b^i D_{x_i} u \overline{v} + cu\overline{v} d\mathbf{x} \\ F(v) &:= \int_\Omega f\overline{v} d\mathbf{x}. \end{aligned}$$

Given these assumptions we use Theorem 3.76 to prove the following result.

Theorem 3.77. *Consider the problem* (3.52) *with all of the restrictions on the coefficients listed above. Let $s \in \mathbb{R}$, with $s \geq 2$ and let $f \in H^{s-2}_p$. Then there exists a unique solution u to* (3.53) *such that $u \in H^s_p$ and*

$$\|u\|_{H^s_p} \leq C\|f\|_{H^{s-2}_p} \tag{3.54}$$

for a constant C.

Proof. Because of our assumptions on the coefficients a^{ij}, b^i and c, the bilinear form $a(\cdot, \cdot)$ is bounded and coercive, and we can apply the Lax-Milgram Lemma to get that (3.53) has a unique solution $u \in H^1_p$ and

$$\|u\|_{H^1_p} \lesssim \|f\|_{H^0_p} \leq \|f\|_{H^{s-2}_p}. \tag{3.55}$$

Define $\theta \in \mathcal{D}(\mathbb{R}^d)$ according to Lemma 3.17 and choose $\Omega' \subset \mathbb{R}^d$ such that $\partial\Omega'$ is of class C^∞ and $\operatorname{supp}\theta \subset \Omega'$ (e.g. choose Ω' to be the open ball with a sufficiently large radius so that $\operatorname{supp}\theta \subset \Omega'$).

Then, by applying L to $w = \theta u$ we see that w is the unique weak solution to the problem

$$\begin{aligned} Lw &= \theta f + g \qquad \text{in } \Omega' \\ w &= 0 \qquad\qquad \text{on } \partial\Omega' \end{aligned}$$

where

$$g = \sum_{i,j=1}^{d} a^{ij}(D_{x_j}\theta)(D_{x_i}u) + a^{ij}(D_{x_i}\theta)(D_{x_j}u) + (D_{x_j}a^{ij}D_{x_i}\theta)u + \sum_{i=1}^{d} b^i(D_{x_i}\theta)u \quad (3.56)$$

Now consider the case when $s = 2$. By Theorem 3.76, $\theta u \in H^2(\Omega')$ and we get

$$\begin{aligned} \|u\|_{H^2_p} &\lesssim \|\theta u\|_{H^2(\mathbb{R}^d)} && \text{by Theorem 3.29} \\ &= \|\theta u\|_{H^2(\Omega')} && \text{since } \operatorname{supp}\theta u \subset \Omega' \\ &\lesssim \|\theta f + g\|_{H^0(\Omega')} && \text{by Theorem 3.76} \\ &\lesssim \|f\|_{H^0_p} + \|g\|_{H^0(\mathbb{R}^d)} && \text{by Theorem 3.29 and extending } g \text{ with zero} \end{aligned}$$

Now choose $\tilde{\theta} \in \mathcal{D}(\mathbb{R}^d)$ to be another function that satisfies the conditions of Lemma 3.17 and define $\tilde{\theta}_{\mathbf{k}}(\mathbf{x}) = \tilde{\theta}(\mathbf{x} + \mathbf{k})$. Since $\operatorname{supp} u \subset \Omega'$, $\tilde{\theta}_{\mathbf{k}} u \neq 0$ for a finite number of $\mathbf{k} \in \mathbb{Z}^d$ and we get the following where the sum is over a finite number of $\mathbf{k} \in \mathbb{Z}^d$,

$$\begin{aligned} \|u\|_{H^2_p} &\lesssim \|f\|_{H^0_p} + \left\| \left(\sum_{\mathbf{k}\in\mathbb{Z}^d} \tilde{\theta}_{\mathbf{k}} \right) g \right\|_{H^0(\mathbb{R}^d)} && \text{since } \textstyle\sum_{\mathbf{k}} \tilde{\theta}_{\mathbf{k}} = 1 \\ &\lesssim \|f\|_{H^0_p} + \sum_{\mathbf{k}\in\mathbb{Z}^d} \left(\sum_{i=1}^{d} \|\tilde{\theta}_{\mathbf{k}}(D_{x_i}u)\|_{H^0(\mathbb{R}^d)} + \|\tilde{\theta}_k u\|_{H^0(\mathbb{R}^d)} \right) && \text{by (3.56)} \end{aligned}$$

where the coefficients of u and D_{x_i} in (3.56) have been absorbed into the constant from "$\lesssim$". Now see that since $\operatorname{supp} u \subset \Omega'$, $\tilde{\theta}_{\mathbf{k}} u \neq 0$ for a finite number of $\mathbf{k} \in \mathbb{Z}^d$ and using Theorem 3.29 we get

$$\begin{aligned} \|u\|_{H^2_p} &\lesssim \|f\|_{H^0_p} + \sum_{i=1}^{d} \|D_{x_i}u\|_{H^0_p} + \|u\|_{H^0_p} \\ &\lesssim \|f\|_{H^0_p} + \|u\|_{H^1_p} \\ &\lesssim \|f\|_{H^0_p} && \text{by (3.55).} \end{aligned}$$

This completes the case $s = 2$.

Now consider the case for general $s \in \mathbb{R}$, $s > 2$. Note first that using (3.54) with

$s = 2$ we also get

$$\|u\|_{H^s_p} \lesssim \|u\|_{H^2_p} \lesssim \|f\|_{H^0_p} \qquad \forall s \in [1,2]. \tag{3.57}$$

Now let $f \in H^{s-2}_p$. By Theorem 3.76, $\theta u \in H^s(\Omega')$ and

$$\begin{aligned}
\|u\|_{H^s_p} &\lesssim \|\theta u\|_{H^s(\mathbb{R}^d)} && \text{by Theorem 3.29} \\
&= \|\theta u\|_{H^s(\Omega')} && \text{since } \operatorname{supp} \theta u \subset \Omega' \\
&\lesssim \|\theta f + g\|_{H^{s-2}(\Omega')} && \text{by Theorem 3.76} \\
&\lesssim \|f\|_{H^{s-2}_p} + \|g\|_{H^{s-2}(\mathbb{R}^d)} && \text{by Theorem 3.29 and extending } g \text{ with zero} \\
&\lesssim \|f\|_{H^{s-2}_p} + \|u\|_{H^{s-1}_p} && \text{by the same argument as for } s = 2
\end{aligned}$$

Now, if $s - 1 \leq 2$ ($s \leq 3$) we can use (3.57) to get

$$\|u\|_{H^s_p} \lesssim \|f\|_{H^{s-2}_p} + \|f\|_{H^0_p} \lesssim \|f\|_{H^{s-2}_p}$$

or, if $s - 1 > 2$ ($s > 3$) we can repeat the arguement above, applying Theorem 3.76 again to get

$$\|u\|_{H^s_p} \lesssim \|f\|_{H^{s-2}_p} + \|f\|_{H^{s-3}_p} + \|u\|_{H^{s-2}_p} \lesssim \|f\|_{H^{s-2}_p} + \|u\|_{H^{s-2}_p}.$$

Now consider whether $s - 2 \leq 2$ and apply (3.57) or apply Theorem 3.76 again. The result follows by repeating the argument above as many times as necessary. □

3.6 Numerical Linear Algebra

In this section we present the tools from numerical linear algebra for solving matrix eigenvalue problems of the form

$$\mathrm{A}\,\mathbf{x} = \lambda \mathbf{x} \qquad \mathrm{A} \in \mathbb{R}^{n \times n}. \tag{3.58}$$

This is not the central focus of this thesis so we will be relatively brief.

We will consider the case when A is symmetric, positive definite (spd) as well as the case when A is unsymmetric. We also note that in practice, we only need to solve (3.58) for the smallest few eigenvalues and corresponding eigenvectors.

The rest of this section is divided into three subsections. In Subsection 3.6.1 we present a Krylov subspace iteration method for finding a subset of the eigenpairs of (3.58). Each step of the method will require us to solve a linear system of the form

$$\mathrm{A}\,\mathbf{x} = \mathbf{b} \tag{3.59}$$

for $\mathbf{x}$ given $\mathbf{b}$. In Subsection 3.6.2 we present the conjugate gradient method (CG)

and the generalised minimal residual method (GMRES) for solving (3.59). Finally, in Subsection 3.6.3 we introduce preconditioning. We rewrite the algorithms for CG and GMRES to include preconditioning and we link the number of iterations required to solve (3.59) to the condition number of the coefficient matrix, where the condition number of a matrix A is defined as

$$\kappa(\mathrm{A}) := \| \mathrm{A} \| \| \mathrm{A}^{-1} \|.$$

Throughout this section we will let MV_c denote the number of operations required to compute a matrix-vector product with A. If A is dense then $MV_c = \mathcal{O}(n^2)$. However, for our numerical examples later in this thesis we will have $MV_c = \mathcal{O}(n \log n)$.

3.6.1 Krylov Subspace Iteration

In this subsection we describe Arnoldi's method for approximating the k most *extremal* eigenvalues of A (i.e k eigenvalues that are away from other eigenvalues). When A is symmetric Arnoldi's method symplifies to Lanczos' method.

The idea of Arnoldi's method is to transform the problem of finding k eigenvalues of A, where A is $n \times n$, to finding k eigenvalues of H, where H is an $m \times m$ upper Hessenberg matrix (only one non-zero sub-diagonal) with $k \leq m \ll n$. The transformation can be achieved through an iterative scheme. A direct method - the QR algorithm - is used to find the eigenvalues of H.

The iterative scheme for transforming A to upper Hessenberg H is called the *Arnoldi process* (not to be confused with the Arnoldi *method*. The Arnoldi method includes computing the eigenvalues and eigenfunctions of H.). We present the Arnoldi process in the following algorithm. Let $\|\cdot\|$ denote here the Euclidean norm for vectors.

Algorithm 3.78. Arnoldi Process. Choose a tolerance $\epsilon_{tol} > 0$ and a starting vector $\mathbf{q}$. The Arnoldi process is as follows:
$\mathbf{q}_1 = \mathbf{q}/\|\mathbf{q}\|$.
For $i = 1, 2, 3, \ldots$
  $\mathbf{v} = \mathrm{A}\,\mathbf{q}_i$
($\star$) For $j = 1, 2, 3, \ldots, i$
    $h_{ji} = \mathbf{q}_j^T \mathbf{v}$
    $\mathbf{v} = \mathbf{v} - h_{ji}\mathbf{q}_j$
  $h_{i+1,i} = \|\mathbf{v}\|$
  If $h_{i+1,i} < \epsilon_{tol}$ and $i \geq k$ then set $\mathbf{q}_{i+1} = \mathbf{v}$, $m = i$ and exit the Arnoldi process.
  If $h_{i+1,i} < \epsilon_{tol}$ and $i < k$ then select random $\mathbf{v}$ and go to ($\star$).
  $\mathbf{q}_{i+1} = \mathbf{v}/h_{i+1,i}$

The output of the Arnoldi process is described by the following lemma which is [73, Propositions 6.5 & 6.6]. The proof of the lemma follows from the algorithm and is

given in [73].

Lemma 3.79. *1. The vectors* $\mathbf{q}_1, \ldots, \mathbf{q}_m$ *form an orthonormal basis for the Krylov subspace* $\mathcal{K}_m = \text{span}\{\mathbf{q}, \mathrm{A}\,\mathbf{q}, \ldots, \mathrm{A}^{m-1}\,\mathbf{q}\}$.

2. If we define a $n \times m$ *matrix* Q_m *with columns* $\mathbf{q}_1, \ldots, \mathbf{q}_m$ *and a* $m \times m$ *matrix* H_m *with entries* h_{ij} *defined by the algorithm then*

$$\begin{aligned} \mathrm{A}\,\mathrm{Q}_m &= \mathrm{Q}_m\,\mathrm{H}_m + \mathbf{q}_{m+1}\mathbf{e}_m^T \\ \mathrm{Q}_m^T\,\mathrm{A}\,\mathrm{Q}_m &= \mathrm{H}_m \end{aligned} \tag{3.60}$$

where $\mathbf{e}_m$ *is an* n*-vector of zeros with a one in the* m^{th} *position.*

The cost of m steps of the Arnoldi process is the cost of m matrix-vector product operations (mMV_c) and the operations to compute H_m and Q_m ($\mathcal{O}(mn)$). Therefore, the total cost of m steps of the Arnoldi process is $\mathcal{O}(mn + mMV_c)$.

The next step of Arnoldi's method is to compute the k largest eigenvalues and corresponding eigenvectors of H_m. This is done using the QR Algorithm. Since H_m is already upper Hessenberg each iteration of the QR Algorithm will cost only $\mathcal{O}(m^2)$ operations since the QR Factorization step will only cost $\mathcal{O}(m^2)$ operations. Assuming that the QR Algorithm converges in $\mathcal{O}(m)$ iterations the total cost of the QR Algorithm will be $\mathcal{O}(m^3)$ operations (see page 194 of [83]).

Therefore, assuming that the Arnoldi process terminates after m steps and that the QR Algorithm converges in $\mathcal{O}(m)$ iterations, then the complete Arnoldi method will cost $\mathcal{O}(mn + mMV_c + m^3)$ operations.

The following theorem explains why the eigenvalues of H_m approximate the eigenvalues of A, thus ensuring that Arnoldi's method works.

Theorem 3.80. *Let* $(\mu, \mathbf{y})$ *be an eigenpair of* H_m *with* $\|\mathbf{y}\| = 1$. *Then* μ *and* $\mathbf{x} := \mathrm{Q}_m\,\mathbf{y}$ *are an approximate eigenpair of* A *with*

$$\|\,\mathrm{A}\,\mathbf{x} - \mu\mathbf{x}\| = h_{m+1,m}|y_m| \leq \epsilon_{tol}$$

where y_m *is the* m^{th} *component of* $\mathbf{y}$.

Proof. From (3.60) we get

$$\begin{aligned} \mathrm{A}\,\mathbf{x} &= \mathrm{A}\,\mathrm{Q}_m\,\mathbf{y} \\ &= \mathrm{Q}_m\,\mathrm{H}_m\,\mathbf{y} + \mathbf{q}_{m+1}\mathbf{e}_m^T\mathbf{y} \\ &= \mathrm{Q}_m\,\mu\mathbf{y} + \mathbf{q}_{m+1}y_m \\ &= \mu\mathbf{x} + \mathbf{q}_{m+1}y_m \end{aligned}$$

The result then follows from

$$\|\mathrm{A}\,\mathbf{x} - \mu\mathbf{x}\| = \|\mathbf{q}_{m+1}\||y_m| = h_{m+1,m}|y_m|.$$

□

It remains to show that the eigenvalues of H_m approximate the extremal eigenvalues of A (i.e. eigenvalues that are away from the other eigenvalues of A). By Theorem 3.80 and Lemma 3.79 we have that the m eigenvectors approximated by the Arnoldi process are in the m-dimensional Krylov subspace $\mathcal{K}_m$. We present a result that estimates the distance between an exact eigenvector of A and $\mathcal{K}_m$. The bound will depend on the initial vector $\mathbf{q}$ and the spectrum of A. A secondary result, will show that the bound is smaller when the exact eigenvector corresponds to an extremal eigenvalue. The following results are from Chapter 6.7 of [73] and we assume that A is diagonalizable.

Theorem 3.81. *Assume that* A *is diagonalizable and that the initial vector* $\mathbf{q}$ *is expanded* $\mathbf{q} = \sum_{j=1}^{n} \alpha_j \mathbf{u}_j$ *with respect to the eigenbasis* $\{\mathbf{u}_j\}_{j=1}^{n}$ *of* A *where* $\|\mathbf{u}_j\|_2 = 1$ *for* $j = 1, \ldots, n$. *Let* P_m *define the orthogonal projection onto* $\mathcal{K}_m$. *Assume that* $\alpha_i \neq 0$ *for some* $i \in \{1, 2, \ldots, n\}$. *Then*

$$\|(I - P_m)u_i\|_2 \leq C_i \epsilon_i^{(m)} \tag{3.61}$$

where

$$C_i = \sum_{\substack{k=1 \\ k \neq i}} \frac{|\alpha_k|}{|\alpha_i|} \qquad \epsilon_i^{(m)} = \min_{\substack{p \in \mathbb{P}_{m-1} \\ p(\lambda_i)=1}} \max_{\substack{\lambda \in \sigma(\mathrm{A}) \\ \lambda \neq \lambda_i}} |p(\lambda)|$$

and $\mathbb{P}_{m-1}$ *denotes the set of all polynomials with degree at most* $m - 1$.

In the Theorem above note that C_i entirely depends the choice of the initial vector $\mathbf{q}$ and that $\mathbf{q}$ must have a component in the direction of the eigenvector we want to approximate. Also note that $\epsilon_i^{(m)}$ only depends on the spectrum of A. We show that Arnoldi's process approximates the extremal eigenvalues of A by showing that $\epsilon_i^{(m)}$ is smaller for λ_i away from other eigenvalues of A. For this we need the following theorem (also from Chapter 6.7 of [73]).

Theorem 3.82. *Let* $m < n$, *let* $i \in \{1, 2, \ldots, n\}$ *and let* (λ_i, u_i) *be an eigenpair of* A. *Then there exist* m *eigenvalues of* A *which can be labelled* $\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,m}$ *such that*

$$\epsilon_i^{(m)} = \left(\sum_{j=1}^{m} \prod_{\substack{k=1 \\ k \neq j}}^{m} \frac{|\lambda_{i,k} - \lambda_i|}{|\lambda_{i,k} - \lambda_{i,j}|} \right)^{-1} \tag{3.62}$$

To bound $\epsilon_i^{(m)}$ from above we should choose $\lambda_{i,1}, \ldots, \lambda_{i,m}$ so that the right-hand-side of (3.62) is as large as possible. This corresponds to choosing $\lambda_{i,1}, \ldots, \lambda_{i,m}$ so that they are relatively as close as possible to λ compared with each other. If λ is far away from the other eigenvalues of A then choosing $\lambda_{i,1}, \ldots, \lambda_{i,m}$ in this way will still give a small upper bound on $\epsilon_i^{(m)}$. However, if λ is clustered together with other eigenvalues then our strategy for choosing $\lambda_{i,1}, \ldots, \lambda_{i,m}$ will result in a large $\epsilon_i^{(m)}$. Therefore, we can construct a smaller bound in (3.61) for an extremal eigenvalue provided our initial guess has a component in the direction of the eigenvector that corresponds to the extremal eigenvalue. This is not a rigorous proof but it agrees with our observations that extremal eigenvalues are approximated first by Arnoldi's method.

As we have stated it, we expect Arnoldi's method to approximate k extremal eigenvalues of a matrix A and these eigenvalues may be the largest or smallest eigenvalues of A (or they may be in the middle of the spectrum if the largest and smallest eigenvalues are densely clustered). If the smallest eigenvalues of A are densely clustered and we want to approximate the smallest k eigenvalues of A then we can apply Arnoldi's method to A^{-1}. The clustered smallest eigenvalues will then become the largest eigenvalues of A^{-1} and they will be (relatively) widely spaced. Similarly, to approximate the k eigenvalues closest to a particular value σ, we replace A in Arnoldi's method with $(\mathrm{A}-\sigma)^{-1}$. Arnoldi's method will then approximate k extremal eigenvalues of $(\mathrm{A}-\sigma)^{-1}$, which we denote by $\mu_1, \ldots, \mu_k$. The k eigenvalues of A closest to σ are then given by $\lambda_i = \frac{1}{\mu_i} + \sigma$ for $i = 1, \ldots, k$. The eigenvector corresponding to μ_i is the eigenvector corresponding to λ_i. We do not necessarily need to store the matrices A^{-1} or $(\mathrm{A}-\sigma)^{-1}$ to calculate these eigenvalues. Since the Arnoldi process only requires the action of A on a vector (the matrix-vector product), we only need the action of A^{-1} or $(\mathrm{A}-\sigma)^{-1}$ on a vector. This can be obtained by solving linear systems of the form of (3.59) or $(\mathrm{A}-\sigma)\mathbf{x} = \mathbf{b}$. This is the topic of the next subsection.

A variation of Arnoldi's method is the Implicitly Restarted Arnoldi Method (IRA) (first published in [77], also described in [87]). The idea of IRA is to reduce the computational cost of Arnoldi's method by limiting the number of steps in the Arnoldi process and therefore limiting the size of the matrices Q_m and H_m. We see from Theorem 3.81 that the convergence of the Arnoldi process depends on the choice of starting vector $\mathbf{q}$. The idea of the IRA method is to restart the Arnoldi process after a fixed number of iterations with a better choice of $\mathbf{q}$, if the Arnoldi process has not already converged. Let $m = \ell + j$ denote when the Arnoldi process will restart. As well as restarting the Arnoldi process, the IRA method also *implicity* computes the first ℓ iterations after each restart. So, after each restart, the IRA method only needs to compute j iterations of the Arnoldi process before the next restart (to *effectively* compute m iterations). The IRA method is not equivalent to Arnoldi's method and some information is lost at each restart.

For the computation of examples in later chapters of this thesis we use the IRA method that is implemented in ARPACK [51].

If A is symmetric then Arnoldi's method becomes Lanczos' method. We replace the Arnoldi process in Algorithm 3.78 with the Lanczos process (see [83] or [13]). The result is an algorithm that computes a symmetric tridiagonal matrix T instead an upper Hessenberg matrix H. The eigenvalues of T then approximate the eigenvalues of A and Theorem 3.80 holds with H replaced with T. The cost of computing the Lanczos process is the same as for the Arnoldi process but the cost of applying the QR algorithm to T is reduced to $\mathcal{O}(m^2)$ operations (from $\mathcal{O}(m^3)$ for Arnoldi) if only eigenvalues are required (assuming that the QR algorithm converges in $\mathcal{O}(m)$ operations). See page 194 of [83] for a discussion of this. Therefore, the total cost of the Lanczos' method is $\mathcal{O}(mn + mMV_c + m^2)$ if only eigenvalues are required. There are more results about the convergence of Lanczos' method to the extremal eigenvalues of A given in [73].

3.6.2 Linear Systems

In this subsection we discuss the problem of solving (3.59) for $\mathbf{x}$ given a right-hand-side $\mathbf{b}$. We present two methods: the conjugate gradient method (CG) and the generalized minimum residual method (GMRES). We use CG when A is symmetric positive definite (spd), otherwise we use GMRES. We begin with CG. The algorithm that follows is from [74].

Algorithm 3.83. CG. Choose a tolerance $\epsilon_{tol} > 0$, a starting vector $\mathbf{x}_0$ and set $\mathbf{r}_0 = \mathbf{p}_0 = \mathbf{b} - \mathrm{A}\,\mathbf{x}_0$.
For $k = 0, 1, 2, \ldots$
  If $\|\mathbf{r}_k\| < \epsilon_{tol}\|\mathbf{r}_0\|$ then exit
  $\alpha = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{p}_k^T \mathrm{A}\,\mathbf{p}_k}$
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \mathbf{p}_k$
  $\mathbf{r}_{k+1} = \mathbf{b} - \alpha\,\mathrm{A}\,\mathbf{p}_k$
  $\beta = \frac{\mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{r}_k^T \mathbf{r}_k}$
  $\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta \mathbf{p}_k$

The CG algorithm has the following two properties that we present in a theorem. These results are Theorems 38.2 and 38.5 of [83]. We omit the proofs.

Theorem 3.84. *Let* A *be spd. Each step of the CG algorithm computes* $\mathbf{x}_k \in \mathbf{x}_0 + \mathcal{K}_k(\mathrm{A}, \mathbf{r}_0)$ *such that* $\|\mathbf{x} - \mathbf{x}_k\|_\mathrm{A}$ *is minimal where* $\mathcal{K}_k(\mathrm{A}, \mathbf{r}_0) = \mathrm{span}\{\mathbf{r}_0, \mathrm{A}\,\mathbf{r}_0, \ldots, \mathrm{A}^{k-1}\,\mathbf{r}_0\}$ *and* $\|\mathbf{y}\|_\mathrm{A} = \sqrt{\mathbf{y}^T\,\mathrm{A}\,\mathbf{y}}$ *is the energy norm induced by* A *(exists for* A *spd). Moreover*

$$\|\mathbf{x} - \mathbf{x}_k\|_\mathrm{A} \leq 2\left(\frac{\sqrt{\kappa(\mathrm{A})} - 1}{\sqrt{\kappa(\mathrm{A})} + 1}\right)^k \|\mathbf{x} - \mathbf{x}_0\|_\mathrm{A}.$$

We can see from this theorem that convergence of the CG method is geometric. However, if $\kappa(\mathrm{A})$ is large then the geometric convergence will be slow. On the other hand, if $\kappa(\mathrm{A})$ is close to one then the convergence of the CG method will be very fast.

Now we discuss GMRES. In some sense it mimics the behaviour of CG for non-symmetric systems, i.e. it is designed to minimise $\|\mathbf{b}-\mathrm{A}\,\mathbf{x}_k\|$ over all $\mathbf{x}_k \in \mathbf{x}_0+\mathcal{K}_k(\mathrm{A},\mathbf{r}_0)$ in some specific norm $\|\cdot\|$. Before we present the GMRES algorithm from page 45 of [43] we must define the following matrices. We define the Given's rotation matrix $G_j(c,s)$ by

$$G_j(c,s) = \begin{bmatrix} 1 & 0 & & \cdots & & & 0 \\ 0 & \ddots & \ddots & & & & \\ & \ddots & c & -s & & & \\ \vdots & & s & c & 0 & & \vdots \\ & & & 0 & 1 & \ddots & \\ & & & & \ddots & \ddots & 0 \\ 0 & & & \cdots & & 0 & 1 \end{bmatrix}$$

where the 2×2 block $\left(\begin{smallmatrix} c & -s \\ s & c \end{smallmatrix}\right)$ is in the j^{th} and $j+1^{st}$ row and column. We also define $\mathrm{Q}_k = G_k(c_k,s_k)\ldots G_1(c_1,s_1)$ and $\mathrm{V}_k = [\mathbf{v}_1\mathbf{v}_2\cdots\mathbf{v}_k]$ where $\mathbf{v}_i$ are orthonormal vectors. We can now define the GMRES algorithm. It is based again on the Arnoldi process.

Algorithm 3.85. GMRES. Choose a tolerance $\epsilon_{tol} > 0$, a maximum number of iterations k_{max}, a starting vector $\mathbf{x}_0$, set $\mathbf{r}_0 = \mathbf{b} - \mathrm{A}\,\mathbf{x}_0$, $\rho = \|\mathbf{r}_0\|$, $\mathbf{v}_1 = \frac{\mathbf{r}_0}{\rho}$ and $\mathbf{g} = \rho\mathbf{e}_1 \in \mathbb{R}^{k_{max}+1}$.
For $k = 1, 2, \ldots, k_{max}$
  If $\rho < \epsilon_{tol}\|\mathbf{b}\|$ then exit
  $\mathbf{v}_{k+1} = \mathrm{A}\,\mathbf{v}_k$
  For $j = 1, \ldots, k$
    $h_{jk} = \mathbf{v}_{k+1}^T\mathbf{v}_j$
    $\mathbf{v}_{k+1} = \mathbf{v}_{k+1} - h_{jk}\mathbf{v}_j$
  $h_{k+1,k} = \|\mathbf{v}_{k+1}\|$
  $\mathbf{v}_{k+1} = \mathbf{v}_{k+1}/h_{k+1,k}$
  If $k > 1$ then apply Q_{k-1} to the k^{th} column of H.
  $\nu = \sqrt{h_{k,k}^2 + h_{k+1,k}^2}$
  $c_k = h_{k,k}/\nu$, $s_k = -h_{k+1,k}/\nu$
  $h_{k,k} = c_k h_{k,k} - s_k h_{k+1,k}$, $h_{k+1,k} = 0$
  $\mathbf{g} = G_k(c_k,s_k)\mathbf{g}$
  $\rho = |\mathbf{g}_{k+1}|$
Set $r_{ij} = h_{ij}$ for $1 \le i,j \le k$
Set $\mathbf{w}_i = \mathbf{g}_i$ for $1 \le i \le k$
Solve upper triangular system $\mathrm{R}\,\mathbf{y}_k = \mathbf{w}$

$\mathbf{x}_k = \mathbf{x}_0 + \mathrm{V}_k\,\mathbf{y}_k$

The cost of each iteration of the GMRES algorithm is $\mathcal{O}(kn + MV_c)$ operations. The break-down of this cost is: MV_c operations for the matrix-vector product; $\mathcal{O}(kn)$ for the orthogonalization procedure (Arnoldi process/Gram-Schmidt); $\mathcal{O}(k^2)$ for the triangular solve; and $\mathcal{O}(kn)$ for constructing $\mathbf{x}_k$. The storage required by the GMRES algorithm is $\mathcal{O}(kn)$ since the $n \times k$ matrix V_k is stored (assuming that we do not store A explicitly).

The GMRES algorithm is also guaranteed to terminate after n iterations (Theorem 3.1.2 on page 34 of [43]). However, if we did in fact iterate up to $k = n$ then GMRES would cost $\mathcal{O}(n^3)$ operations and the storage requirement would be $\mathcal{O}(n^2)$.

Often it is the storage requirement that makes standard GMRES impractical. To alleviate the storage requirements of GMRES we use a variation of GMRES: Restarted GMRES. In Restarted GMRES we set $k_{max} = m \ll n$ and restart the algorithm with $\mathbf{x}_0 = \mathbf{x}_m$ if it does not terminate before $k = k_{max}$. Restarted GMRES is not equivalent to GMRES because the information in V_m (the basis for $\mathcal{K}(\mathrm{A}, \mathbf{r}_0)$ is discarded when the algorithm is restarted. For this reason, [43, Theorem 3.1.2] can not be applied to Restarted GMRES and it is not guaranteed to terminate. However, it works well in practice and only requires $\mathcal{O}(mn)$ storage.

The residual at each iteration of GMRES can be bounded in the following way.

Theorem 3.86. *At each step k of GMRES, the residual $\mathbf{r}_k$ is bounded by*

$$\frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|} \leq \inf_{p_k \in \mathcal{P}_k} \|p_k(\mathrm{A})\|$$

where $\mathcal{P}_k$ is the space of all degree k polynomials. If A is diagonalizable we may write $\mathrm{A} = \mathrm{V}\,\Lambda\,\mathrm{V}^{-1}$ where V is orthogonal and Λ is diagonal containing the eigenvalues of A. Then

$$\frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_0\|} \leq \kappa(\mathrm{V}) \inf_{p_k \in \mathcal{P}_k} \sup_{\lambda \in \Lambda(\mathrm{A})} |p_k(\lambda)|$$

where $\Lambda(\mathrm{A})$ is the set of all eigenvalues of A.

In Theorem 3.84 we saw that the convergence of CG depended on $\kappa(\mathrm{A}) = \frac{\lambda_{max}}{\lambda_{min}}$ (for $\|\cdot\| = \|\cdot\|_2$), i.e. the convergence of CG depends only on the spectrum of A. This is in contrast to GMRES where in Theorem 3.86 we see that the convergence depends on the eigenfunctions of A (through $\kappa(\mathrm{V})$) as well as the spectrum of A.

3.6.3 Preconditioning Linear Systems

In this subsection we discuss the technique called *preconditioning* that is used to make (3.59) easier to solve. Instead of solving (3.59), we solve

$$(\mathrm{P}^{-1}\,\mathrm{A})\mathbf{x} = (\mathrm{P}^{-1}\,\mathbf{b}) \tag{3.63}$$

where the matrix P is called the preconditioner. The idea is to choose P so that the condition number of $\mathrm{P}^{-1}\,\mathrm{A}$ is less than the condition number of A and the operation of P^{-1} cheap to compute.

In both the CG method and the GMRES method we have chosen to terminate when the relative residual $\frac{\|\mathbf{b}-\mathrm{A}\,\mathbf{x}_k\|}{\|\mathbf{b}-\mathrm{A}\,\mathbf{x}_0\|}$ is bounded by a tolerance ϵ_{tol}. We hope that the relative residual gives a good indication of the actual error in $\mathbf{x}_k$. It is possible to derive the following bound, where $\mathbf{x}^\star$ is the exact solution to (3.59),

$$\frac{\|\mathbf{x}_k-\mathbf{x}^\star\|}{\|\mathbf{x}_0-\mathbf{x}^\star\|}\leq\kappa(\mathrm{A})\frac{\|\mathbf{b}-\mathrm{A}\,\mathbf{x}_k\|}{\|\mathbf{b}-\mathrm{A}\,\mathbf{x}_0\|}.$$

Therefore, if we choose P so that $\kappa(\mathrm{P}^{-1}\,\mathrm{A})\ll\kappa(\mathrm{A})$ then both CG and GMRES will terminate when the relative residual error is a more acurate bound of the actual relative error.

As well as achieving a better indication of the actual error by preconditioning (3.59) we also achieve faster convergence for either CG or GMRES through preconditioning. First, we will present the Preconditioned Conjugate Gradient (PCG) method, then we consider preconditioning with GMRES.

For the CG method the coefficient matrix must be spd but for A and P^{-1} spd, $\mathrm{P}^{-1}\,\mathrm{A}$ is unsymmetric in general. Therefore, if we want to solve the preconditioned linear system we must choose P^{-1} spd and solve

$$(\mathrm{P}^{-1/2}\,\mathrm{A}\,\mathrm{P}^{-1/2})\mathbf{y}=\mathrm{P}^{-1/2}\,\mathbf{b}\tag{3.64}$$

for $\mathbf{y}$. The solution of (3.59) is then given by $\mathbf{x}=\mathrm{P}^{-1/2}\,\mathbf{y}$. The following algorithm is called the PCG method and solves (3.64) without having to apply or calculate $\mathrm{P}^{-1/2}$. It is from page 246 of [74]. It is just Algorithm 3.83 constructed with the P^{-1} norm and inner product, $\|\mathbf{x}\|_{P^{-1}}=\sqrt{\mathbf{x}^T\,\mathrm{P}^{-1}\,\mathbf{x}}$, $(\mathbf{x},\mathbf{y})_{P^{-1}}=\mathbf{x}^T\,\mathrm{P}^{-1}\,\mathbf{y}$.

Algorithm 3.87. PCG. Choose a tolerance $\epsilon_{tol}>0$, starting vector $\mathbf{x}_0$, set $\mathbf{r}_0=\mathbf{b}-\mathrm{A}\,\mathbf{x}_0$ and $\mathbf{z}_0=\mathbf{p}_0=\mathrm{P}^{-1}\,\mathbf{r}_0$.
For $k=0,1,2,\ldots$
If $\|\mathbf{r}_k\|<\epsilon_{tol}\|\mathbf{r}_0\|$ then exit
$\alpha=\frac{\mathbf{z}_k^T\,\mathbf{r}_k}{\mathbf{p}_k^T\,\mathrm{A}\,\mathbf{p}_k}$
$\mathbf{x}_{k+1}=\mathbf{x}_k+\alpha\mathbf{p}_k$
$\mathbf{r}_{k+1}=\mathbf{b}-\alpha\,\mathrm{A}\,\mathbf{p}_k$
$\mathbf{z}_{k+1}=\mathrm{P}^{-1}\,\mathbf{r}_{k+1}$
$\beta=\frac{\mathbf{z}_{k+1}^T\mathbf{r}_{k+1}}{\mathbf{z}_k^T\mathbf{r}_k}$
$\mathbf{p}_{k+1}=\mathbf{z}_{k+1}+\beta\mathbf{p}_k$

If $\kappa(\mathrm{P}^{-1}\,\mathrm{A})\ll\kappa(\mathrm{A})$ then Theorem 3.84 guarantees that using the PCG method will converge faster than the CG method.

In the case of the GMRES method, we do not require that the coefficient matrix is symmetric or positive definite. Therefore, we are free to choose P^{-1} without restriction and we simply apply GMRES to (3.63). Algorithm 3.85 must be modified in two steps. In the initial set up we compute the initial residual as $\mathbf{r}_0 = \mathrm{P}^{-1}(\mathbf{b} - \mathrm{A}\,\mathbf{x}_0)$ and we replace the step $\mathbf{v}_{k+1} = \mathrm{A}\,\mathbf{v}_k$ with $\mathbf{v}_{k+1} = \mathrm{P}^{-1}\,\mathrm{A}\,\mathbf{v}_k$.

Theorem 3.86 implies that to choose a good preconditioner for GMRES we should choose P so that

$$\inf_{p_k \in \mathcal{P}_k} \|p_k(\mathrm{P}^{-1}\,\mathrm{A})\| \ll \inf_{p_k \in \mathcal{P}_k} \|p_k(\mathrm{A})\|.$$

If $\mathrm{P}^{-1}\,\mathrm{A}$ is diagonalizable and $\mathrm{P}^{-1}\,\mathrm{A} = \mathrm{V}\,\Lambda\,\mathrm{V}^{-1}$ then we want to have chosen P^{-1} so that $\kappa(\mathrm{V})$ is small and

$$\inf_{p_k \in \mathcal{P}_k} \sup_{\lambda \in \Lambda(\mathrm{P}^{-1}\,\mathrm{A})} |p_k(\lambda)| \ll \inf_{p_k \in \mathcal{P}_k} \sup_{\lambda \in \Lambda(\mathrm{A})} |p_k(\lambda)|.$$

CHAPTER 4

SCALAR 2D PROBLEM & 1D TE MODE PROBLEM

In this chapter we solve the Scalar 2D Problem (2.19) and the 1D TE Mode Problem (2.20), as defined in Chapter 2, using the plane wave expansion method, and variations of the plane wave expansion method. As well as including details for the efficient implementation of the different methods, the main emphasis of this chapter will be on the error convergence analysis for each of the methods.

Since the 1D and 2D problems are very similar we will focus on the 2D problem most of the time. Indeed, we will find that the same theory applies to both problems more often than not, but where there are differences between the problems we will point these out.

The chapter is divided into six sections. In the first section we introduce the 2D problem as an operator on a Hilbert space with unknown spectrum that we would like to approximate. We apply the Floquet Transform from Subsection 3.4.3 to obtain a family of operators on a bounded domain, each with discrete spectrum. Therefore, we can write down a variational eigenvalue problem corresponding to each new operator. We then prove a regularity result for the eigenfunctions of the variational problems. Next, we consider the special features of the 1D problem before defining examples that will be referred to throughout this chapter.

In the second section we apply the plane wave expansion method to the variational eigenvalue problem. We then include implementation details for the method before we develop a full error convergence analysis for the standard plane wave expansion method.

In Sections 4.3 - 4.5 we present variations of the plane wave expansion method: the smoothing method, the sampling method, and the smoothing and sampling method. We include implementation details together with error convergence analysis for each of

these methods.

In the final section we briefly discuss an expansion method based on curvilinear coordinates and how we lose an optimal preconditioner for this method.

Throughout this chapter we will make use of the mathematical tools that we presented in Chapter 3.

4.1 The Problem

4.1.1 The Spectral Problem

From (2.19) the formal equation for the Scalar 2D Problem is

$$\nabla^2 h + \gamma(\mathbf{x})h = \beta^2 h \tag{4.1}$$

where ∇ is the 2D gradient operator, $h = h(\mathbf{x})$ is a 2D scalar field, β^2 is an eigenvalue and $\gamma(\mathbf{x})$ is a 2D scalar field that is periodic on a Bravais lattice in $\mathbb{R}^2$. For simplicity and as discussed in Section 3.2, we restrict all of our presentation to the Bravais lattice $\mathbb{Z}^2$ with period cell $\Omega = (-\frac{1}{2}, \frac{1}{2})^2$. We also assume that $\gamma \in PC_p$, i.e. $\gamma(\mathbf{x})$ is in our special class of piecewise continuous functions that we defined in Definition 3.36. This implies that $\gamma \in L_p^\infty$ and without loss of generality we specify that $0 < \gamma(\mathbf{x}) \leq \gamma_{\max}$ for all $\mathbf{x} \in \mathbb{R}^2$. For some results we will also assume certain symmetries of $\gamma(\mathbf{x})$ or that $\gamma \in PC_p'$ (see Definition 3.37).

The aim is to find the unknown eigenvalues β^2 and the corresponding eigenfunctions h of (4.1).

Mathematically, we state our problem as a spectral problem. We want to find the spectrum of an operator on a Hilbert space. For this problem the Hilbert space is $L^2(\mathbb{R}^2)$ with the usual inner product and the operator is

$$\mathrm{L} := -\nabla^2 - \gamma(\mathbf{x}) + \mathrm{K} \tag{4.2}$$

with domain $H^2(\mathbb{R}^2)$. To obtain L from (4.1) we have multiplied (4.1) by -1 and we have added a constant K to shift the spectrum and ensure that L is always positive definite. If $\lambda \in \sigma(\mathrm{L})$ then we say that $\beta^2 = -\lambda + \mathrm{K}$ is an eigenvalue of (4.1). For now, we will only say that K is sufficiently large to ensure that L is positive definite. We will be more specific about our choice of K later.

The following result is a well known classical result.

Theorem 4.1. *The spectrum of* L *is real and purely essential, i.e.*

$$\sigma(\mathrm{L}) = \sigma_{ess}(\mathrm{L}) \subset \mathbb{R}.$$

where $\sigma_{ess}(\mathrm{L})$ *denotes the essential spectrum of* L.

Proof. It is easy to see that L is self-adjoint and $\sigma(\mathrm{L}) \subset \mathbb{R}$ follows from Theorem 3.60. $\sigma(\mathrm{L}) = \sigma_{ess}(\mathrm{L})$ follows from Theorem XIII.100 on page 309 of [69]. □

We are only interested in the spectrum of (4.1) that lies in the positive real half plane. This is because eigenvalues with a negative real part correspond to *evanescent* eigenfunctions (i.e. non-physical electromagnetic waves). Therefore, we are only interested in the spectrum of L that is in the interval $[0, \mathrm{K}]$.

4.1.2 Applying the Floquet Transform

Since the coefficients of L are periodic we can apply the Floquet Transform to L on $L^2(\mathbb{R}^2)$ as in [17], [45] or [78]. We defined the Floquet Transform in Subsection 3.4.3. After applying the transform we obtain a family of operators parameterised by $\boldsymbol{\xi} \in B$ on a bounded domain, where $B = [-\pi, \pi]^2$ is the 1st Brillouin Zone corresponding to the Bravais lattice $\mathbb{Z}^2$. In photonics literature $\boldsymbol{\xi}$ is called the *quasi-momentum*. The transformed problems are posed on the Hilbert space L^2_p with the usual $L^2(\Omega)$ inner product, where Ω is the period cell of the Bravais lattice. For each $\boldsymbol{\xi} \in B$ the operator is defined as

$$\mathrm{L}_{\boldsymbol{\xi}} := -(\nabla + i\boldsymbol{\xi})^2 - \gamma(\mathbf{x}) + \mathrm{K}$$

with domain H^2_p (defined in Section 3.2). We can prove the following properties about the spectrum of our new family of operators.

Lemma 4.2. *The spectrum of* $\mathrm{L}_{\boldsymbol{\xi}}$ *has the following properties:*

1. $\sigma(\mathrm{L}_{\boldsymbol{\xi}}) \subset \mathbb{R}$ *for every* $\boldsymbol{\xi} \in B$.
2. $\sigma(\mathrm{L}_{\boldsymbol{\xi}}) = \sigma_d(\mathrm{L}_{\boldsymbol{\xi}}))$ *where* $\sigma_d(\mathrm{L}_{\boldsymbol{\xi}})$ *denotes the discrete spectrum of* $\mathrm{L}_{\boldsymbol{\xi}}$ *for every* $\boldsymbol{\xi} \in B$.
3. $\lambda(\boldsymbol{\xi}) \in \sigma(\mathrm{L}_{\boldsymbol{\xi}})$ *considered as a function of* $\boldsymbol{\xi}$ *is continuous on* B.

Proof. 1. $\sigma(\mathrm{L}_{\boldsymbol{\xi}}) \subset \mathbb{R}$ follows from the fact that $\mathrm{L}_{\boldsymbol{\xi}}$ is self-adjoint with domain $\mathcal{D}(\mathrm{L}_{\boldsymbol{\xi}}) = H^2_p$. To see that $\mathrm{L}_{\boldsymbol{\xi}}$ is self-adjoint, notice that we have $(\mathrm{L}_{\boldsymbol{\xi}}\, u, v)_{L^2(\Omega)} = (u, \mathrm{L}_{\boldsymbol{\xi}}\, v)_{L^2(\Omega)}$ for all $u, v \in \mathcal{D}(\mathrm{L}_{\boldsymbol{\xi}})$, using integration by parts. This implies that $\mathrm{L}_{\boldsymbol{\xi}}$ is symmetric, i.e. $\mathcal{D}(\mathrm{L}_{\boldsymbol{\xi}}) \subset \mathcal{D}(\mathrm{L}^*_{\boldsymbol{\xi}})$. Moreover, in the above working for the integration by parts we require that $v \in H^2_p(\Omega)$ for $(\mathrm{L}_{\boldsymbol{\xi}}\, u, v)_{L^2(\Omega)} = (u, \mathrm{L}_{\boldsymbol{\xi}}\, v)_{L^2(\Omega)}$ to hold for all $u \in \mathcal{D}(\mathrm{L}_{\boldsymbol{\xi}})$. Therefore, $\mathcal{D}(\mathrm{L}^*_{\boldsymbol{\xi}}) = \mathcal{D}(\mathrm{L}_{\boldsymbol{\xi}})$ and $\mathrm{L}_{\boldsymbol{\xi}} = \mathrm{L}^*_{\boldsymbol{\xi}}$.

2. According to part a) of Lemma 2 on page 308 of [69] there exists a $\mu \notin \sigma(\mathrm{L}_{\boldsymbol{\xi}})$ such that the resolvant of $\mathrm{L}_{\boldsymbol{\xi}}$ is compact. Therefore, the spectrum of $\mathrm{L}_{\boldsymbol{\xi}}$ is purely discrete by part 3 of Theorem 3.60.

3. This result follows from the discussion in [69] and is stated in [69, Lemma 2 on page 308]. □

Part 2 of Lemma 4.2 is a useful result for developing a numerical method because a numerical method will attempt to approximate $\mathrm{L}_{\boldsymbol{\xi}}$ with an operator on a finite dimensional Hilbert space and such an operator will also have discrete spectrum. If $\mathrm{L}_{\boldsymbol{\xi}}$ had essential spectrum then it would be difficult to measure the accuracy of our numerical method because it is not clear how the discrete spectrum from an approximate problem would approximate essential spectrum of $\mathrm{L}_{\boldsymbol{\xi}}$.

Part 3 of Lemma 4.2 is also a useful result in light of the next theorem as it tells us how the discrete spectrum of $\mathrm{L}_{\boldsymbol{\xi}}$ will approximate the essential spectrum of L. We can take advantage of the contininuity of the eigenvalues with respect to $\boldsymbol{\xi}$ by only approximating the spectrum of $\mathrm{L}_{\boldsymbol{\xi}}$ for a finite number of $\boldsymbol{\xi} \in B$.

Now we apply the key result from Floquet theory, Theorem 3.63, to get the following result.

Theorem 4.3.

$$\sigma(\mathrm{L}) = \bigcup_{\boldsymbol{\xi} \in B} \sigma(\mathrm{L}_{\boldsymbol{\xi}})$$

If $\gamma(\mathbf{x})$ has certain symmetries, then we get the following result. This type of result can also be found in [39].

Corollary 4.4. *If $\gamma \in PC_p$ and $\gamma(x_1, x_2) = \gamma(-x_1, x_2) = \gamma(x_1, -x_2) = \gamma(x_2, x_1)$ for all $x_1, x_2 \in \mathbb{R}$. Then $\lambda(\boldsymbol{\xi}) \in \sigma(\mathrm{L}_{\boldsymbol{\xi}})$ also has these symmetries for all $\boldsymbol{\xi} \in B$, i.e. $\lambda(\xi_1, \xi_2) = \lambda(-\xi_1, \xi_2) = \lambda(\xi_1, -\xi_2) = \lambda(\xi_2, \xi_1)$ for all $\xi_1, \xi_2 \in \mathbb{R}$ and*

$$\sigma(\mathrm{L}) = \bigcup_{\boldsymbol{\xi} \in B_I} \sigma(\mathrm{L}_{\boldsymbol{\xi}})$$

where B_I is the irreducible Brillouin zone *defined as the triangular region with vertices $(0,0)$, $(\pi, 0)$, and (π, π), i.e. $B_I = \{\boldsymbol{\xi} \in B : 0 \leq \xi_1 \leq \pi, 0 \leq \xi_2 \leq \xi_1\}$.*

Proof. We will prove that if $\gamma(x_1, x_2) = \gamma(-x_1, x_2)$ for all $x_1, x_2 \in \mathbb{R}$ then $\lambda(\xi_1, \xi_2) = \lambda(-\xi_1, \xi_2)$ for all $\xi_1, \xi_2 \in \mathbb{R}$ for all $\lambda(\boldsymbol{\xi}) \in \sigma(\mathrm{L}_{\boldsymbol{\xi}})$. The results for mirror symmetries in the other directions are proved in a similar way.

Let $\boldsymbol{\xi} \in B$ and let $\mathbf{y}(\mathbf{x}) = (-x_1, x_2)^T$. By Part 2 of Lemma 4.2 we know that the spectrum of $\mathrm{L}_{\boldsymbol{\xi}}$ is discrete. Therefore any $\lambda(\boldsymbol{\xi}) \in \sigma(\mathrm{L}_{\boldsymbol{\xi}})$ is an eigenvalue of $\mathrm{L}_{\boldsymbol{\xi}}$ with corresponding eigenfunction $u(\mathbf{x})$. We will show that $(\lambda(\boldsymbol{\xi}), u(\mathbf{y}(\mathbf{x}))$ is an eigenpair of $\mathrm{L}_{(-\xi_1, \xi_2)}$. Using the chain rule we get

$$\nabla_{\mathbf{x}} u(\mathbf{y}(\mathbf{x})) = \begin{pmatrix} \frac{\partial u}{\partial y_1}\frac{\partial y_1}{\partial x_1} + \frac{\partial u}{\partial y_2}\frac{\partial y_2}{\partial x_1} \\ \frac{\partial u}{\partial y_1}\frac{\partial y_1}{\partial x_2} + \frac{\partial u}{\partial y_2}\frac{\partial y_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} -\frac{\partial u}{\partial y_1} \\ \frac{\partial u}{\partial y_2} \end{pmatrix}$$

And so,

$$\begin{aligned}
\mathrm{L}_{(-\xi_1,\xi_2)}\, u(\mathbf{y}(\mathbf{x})) &= \left(\begin{pmatrix} -\frac{\partial u}{\partial \mathbf{y}_1} \\ \frac{\partial u}{\partial \mathbf{y}_2} \end{pmatrix} + i \begin{pmatrix} -\xi_1 \\ \xi_2 \end{pmatrix} \right)^2 u(\mathbf{y}(\mathbf{x})) - \gamma(\mathbf{x}) u(\mathbf{y}(\mathbf{x})) + \mathrm{K}\, u(\mathbf{y}(\mathbf{x})) \\
&= (\nabla_{\mathbf{y}} + i\boldsymbol{\xi})^2 u(\mathbf{y}) - \gamma(\mathbf{y}) u(\mathbf{y}) + \mathrm{K}\, u(\mathbf{y}) \\
&= \mathrm{L}_{\boldsymbol{\xi}}\, u(\mathbf{y}) \\
&= \lambda(\boldsymbol{\xi}) u(\mathbf{y})
\end{aligned}$$

and so $(\lambda(\boldsymbol{\xi}), u(\mathbf{y}(\mathbf{x})))$ is an eigenpair of $\mathrm{L}_{(-\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)}$. It follows that $\lambda(\boldsymbol{\xi})$ as a function of $\boldsymbol{\xi} \in B$ is mirror symmetric with respect to the ξ_1 coordinate direction.

The final statement that $\sigma(\mathrm{L}) = \bigcup_{\boldsymbol{\xi} \in B_I} \sigma(\mathrm{L}_{\boldsymbol{\xi}})$ follows from Theorem 4.3 and the symmetries of $\lambda(\boldsymbol{\xi}) \in \sigma(\mathrm{L}_{\boldsymbol{\xi}})$. □

Finally, we state an unproven conjecture that is often used (implicitly) in photonics literature, see for example, [5], [8], [15], [34], [38] and [79].

The conjecture allows us to make a further restriction on the $\boldsymbol{\xi} \in B$ that we need to consider. If the conjecture holds then we only need to consider $\boldsymbol{\xi} \in \partial B_I$ where ∂B_I is the boundary of B_I (B_I is defined in Corollary 4.4).

Conjecture 4.5. *Assume that $\gamma \in PC_p$ satisfies the symmetries in Corollary 4.4. For any $\boldsymbol{\xi} \in B$ let $\lambda_j(\boldsymbol{\xi})$ denote the the j^{th} smallest eigenvalue in $\sigma(\mathrm{L}_{\boldsymbol{\xi}})$. Define*

$$\lambda_{j,min} := \min_{\boldsymbol{\xi}' \in \partial B_I} \lambda_j(\boldsymbol{\xi}') \qquad \lambda_{j,max} := \max_{\boldsymbol{\xi}' \in \partial B_I} \lambda_j(\boldsymbol{\xi}')$$

where ∂B_I is the boundary of B_I (B_I is defined in Corollary 4.4). Then

$$\lambda_j(\boldsymbol{\xi}) \in [\lambda_{j,min}, \lambda_{j,max}]\,.$$

The significant consequence of this conjecture is that to approximate $\sigma(\mathrm{L})$ we only need to compute $\sigma(\mathrm{L}_{\boldsymbol{\xi}})$ for $\xi \in \partial B_I$, i.e. we only need to compute the spectrum of $\mathrm{L}_{\boldsymbol{\xi}}$ on the *boundary* of the irreducible Brillouin zone. This is a significant saving in computational cost because without the conjecture we would need to compute $\sigma(\mathrm{L}_{\boldsymbol{\xi}})$ for all $\xi \in B_I$.

An alternative approach that is sometimes used in the photonics literature (see for example [66]), that does not rely on this conjecture, is the *density of states* method. The *density of states* method samples $\xi \in B_I$ (usually on a uniform grid) and counts the number of times that an eigenvalue appears in a small interval of possible β^2 values and in a small frequency range. This count determines the *density* of the state where the *state* is determined by the small range of frequencies and small range of β^2. A plot is then drawn for the density of states vs. both frequency and β^2. Regions where the density of states is low are considered to be bandgaps.

For the 1D problem, defined later in this section, Conjecture 4.5 has been proven and the result can be found on page 293 of [69]. We present an equivalent result in Lemma 4.14.

In this thesis we will rely on Conjecture 4.5. Let us now focus on the central problem of approximating the spectrum of $\mathrm{L}_{\boldsymbol{\xi}}$ for a fixed $\boldsymbol{\xi} \in B$.

4.1.3 Variational Formulation

In this subsection we take advantage of the fact that the spectrum of $\mathrm{L}_{\boldsymbol{\xi}}$ is discrete and we write down the variational eigenvalue problem which, under additional regularity assumptions, is equivalent to finding a $\lambda \in \sigma(\mathrm{L}_{\boldsymbol{\xi}})$ and its corresponding eigenfunction. The variational eigenvalue problem is defined as

Problem 4.6. For a fixed $\boldsymbol{\xi} \in B$, find an eigenpair (λ, u) where $\lambda \in \mathbb{C}$ and $0 \neq u \in H^1_p$ such that

$$a(u,v) = \lambda b(u,v) \qquad \forall v \in H^1_p \tag{4.3}$$

where

$$\begin{aligned} a(u,v) &= \int_\Omega (\nabla + i\boldsymbol{\xi})\, u \cdot \overline{(\nabla + i\boldsymbol{\xi})\, v} + (\mathrm{K} - \gamma)\, u\overline{v} dx \\ b(u,v) &= \int_\Omega u\overline{v} dx. \end{aligned}$$

We will now prove some properties of the bilinear form $a(\cdot,\cdot)$ that will enable us to say more about the spectral properties of Problem 4.6. The following lemma will also be very important for the error convergence results later in this chapter.

Lemma 4.7. *Provided we choose* $\mathrm{K} \geq \gamma_{\max} + 2\pi^2 + \frac{1}{2}$*, the bilinear form* $a(\cdot,\cdot)$ *from Problem 4.6 is bounded, coercive and Hermitian on* H^1_p*.*

Proof. Part 1. $a(\cdot,\cdot)$ bounded.

$$\begin{aligned} |a(u,v)| &= \left| \int_\Omega (\nabla + i\boldsymbol{\xi})\, u \cdot \overline{(\nabla + i\boldsymbol{\xi})\, v} + (\mathrm{K} - \gamma)\, u\overline{v} dx \right| \\ &\leq \int_\Omega |\nabla u \cdot \overline{\nabla v} + i\boldsymbol{\xi} u \cdot \overline{\nabla v} - i\boldsymbol{\xi} \cdot \nabla u \overline{v} + (|\boldsymbol{\xi}|^2 + \mathrm{K} - \gamma) u\overline{v}| dx \\ &\leq \left(1 + 2|\boldsymbol{\xi}| + |\boldsymbol{\xi}|^2 + \|\,\mathrm{K} - \gamma\|_\infty\right) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \\ &\leq \left((1+\pi)^2 + \mathrm{K}\right) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \\ &= C \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \\ &= C \|u\|_{H^1_p} \|v\|_{H^1_p} \qquad \forall u, v \in H^1_p \end{aligned}$$

with $C = (1+\pi)^2 + \mathrm{K}$.

Part 2. $a(\cdot,\cdot)$ coercive. We will use the Cauchy Schwarz inequality (CS) and the arithmetic-geometric mean inequality (AG), that says $2xy \leq x^2 + y^2$.

$$\begin{aligned}
a(v,v) &= \int_\Omega (\nabla + i\boldsymbol{\xi})\, v \cdot \overline{(\nabla + i\boldsymbol{\xi})\, v} + (\mathrm{K} - \gamma)\, v\overline{v} dx \\
&= \int_\Omega |\nabla v|^2 + i\boldsymbol{\xi} v \cdot \overline{\nabla v} - i\boldsymbol{\xi} \cdot \nabla v \overline{v} + (|\boldsymbol{\xi}|^2 + \mathrm{K} - \gamma)|v|^2 dx \\
&\geq |v|^2_{H^1(\Omega)} + (|\boldsymbol{\xi}|^2 + \mathrm{K} - \gamma_{\max})\|v\|^2_{L^2(\Omega)} - 2|\boldsymbol{\xi}||v|_{H^1}\|v\|_{L^2(\Omega)} \qquad \text{(CS)} \\
&\geq |v|^2_{H^1(\Omega)} + (|\boldsymbol{\xi}|^2 + \mathrm{K} - \gamma_{\max})\|v\|^2_{L^2(\Omega)} - \tfrac{1}{2}\left(|v|^2_{H^1(\Omega)} + 4|\boldsymbol{\xi}|^2\|v\|^2_{L^2(\Omega)}\right) \qquad \text{(AG)} \\
&= \tfrac{1}{2}|v|^2_{H^1(\Omega)} + (-|\boldsymbol{\xi}|^2 + \mathrm{K} - \gamma_{\max})\|v\|^2_{L^2(\Omega)} \\
&\geq \tfrac{1}{2}|v|^2_{H^1(\Omega)} + (-2\pi^2 + \mathrm{K} - \gamma_{\max})\|v\|^2_{L^2(\Omega)} \\
&\geq C\|v\|^2_{H^1(\Omega)} \\
&= C\|v\|^2_{H^1_p} \qquad \forall v \in H^1_p
\end{aligned}$$

with $C = \frac{1}{2}$ provided that K is chosen so that $\mathrm{K} \geq \gamma_{\max} + 2\pi^2 + \frac{1}{2}$.

Part 3. $a(\cdot,\cdot)$ Hermitian. The proof that $a(\cdot,\cdot)$ is Hermitian is obvious from the definition of $a(\cdot,\cdot)$. □

Also note that $b(\cdot,\cdot)$ from Problem 4.6 is the usual $L^2(\Omega)$ inner product and it is bounded and Hermitian on L^2_p.

The previous lemma leads directly to the following corollary that will be necessary for later in the chapter.

Corollary 4.8. *$a(\cdot,\cdot)$ defines an inner product on H^1_p and the induced norm $\|\cdot\|_a = a(\cdot,\cdot)^{\frac{1}{2}}$ is equivalent to $\|\cdot\|_{H^1_p}$.*

4.1.4 Properties of the Spectrum

In this subsection we introduce the *solution operator* corresponding to Problem 4.6 as a means of proving more results about the spectrum of $\mathrm{L}_{\boldsymbol{\xi}}$. We will also use the solution operator later in the chapter as a tool for proving error convergence results.

The solution operator T corresponding to Problem 4.6 is defined according to Definition 3.70 in Subsection 3.5.2 with $\mathcal{H} := H^1_p$. The following lemma proves some basic properties of T.

Lemma 4.9. *The solution operator* T *corresponding to Problem 4.6 has the following properties*

1. $\mathrm{T} : L^2_p(\Omega) \to H^1_p(\Omega)$ *is bounded.*
2. $\mathrm{T} : H^1_p \to H^1_p$ *is compact.*
3. $\mathrm{T} : H^1_p \to H^1_p$ *is self-adjoint with respect to $a(\cdot,\cdot)$.*

4. $\mathrm{T} : H^1_p \to H^1_p$ *is positive definite with respect to* $a(\cdot,\cdot)$.

Proof. Part 1. $\mathrm{T} : L^2_p \to H^1_p$ bounded follows from the Lax-Milgram Lemma since $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are bounded and $a(\cdot,\cdot)$ is coercive (Lemma 4.7).

Part 2. Since H^1_p is compactly embedded in L^2_p (Theorem 3.24), the inclusion operator $I : H^1_p \to L^2_p$ is compact. Using this with Part 1 it follows that $\mathrm{T} : H^1_p \to H^1_p$ is compact (since the composition of a compact operator and a linear bounded operator is compact, see page 233-234 of [50]). Using a similar argument we can also show that $\mathrm{T} : L^2_p \to L^2_p$ is compact.

Part 3. $\mathrm{T} : H^1_p \to H^1_p$ is symmetric with respect to $a(\cdot,\cdot)$ since $a(\mathrm{T}f, g) = b(f,g) = \overline{b(g,f)} = \overline{a(\mathrm{T}g, f)} = a(f, \mathrm{T}g)$ for all $f, g \in H^1_p$. $\mathrm{T} : H^1_p \to H^1_p$ is also bounded with respect to $\|\cdot\|_a$ (the norm induced by $a(\cdot,\cdot)$). Therefore, $\mathrm{T} : H^1_p \to H^1_p$ is self-adjoint with respect to $a(\cdot,\cdot)$.

Part 4. $a(\mathrm{T}f, f) = b(f,f) > 0$ for all $0 \neq f \in H^1_p$. □

Now we use these properties of the solution operator to describe the spectrum of Problem 4.6. Before we write down the result and proof, note that since T is compact and self-adjoint on a Hilbert space we know that the ascent of any eigenvalue of T will be 1 and algebraic multiplicity is equal to geometric multiplicity. Therefore, we do not need to consider generalised eigenfunctions. See our comments in Subsection 3.4.2. This reasoning is also used on page 683 of [6].

Lemma 4.10. *Problem 4.6 has eigenvalues*

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \nearrow +\infty$$

counted up to multiplicity (i.e. if λ_j *has multiplicity* 2 *then set* $\lambda_{j+1} = \lambda_j$*) with corresponding eigenfunctions*

$$u_1, u_2, \ldots$$

that can be chosen such that

$$a(u_i, u_j) = \delta_{ij} \qquad \forall i, j \in \mathbb{N}.$$

Moreover, the eigenfunctions are complete in L^2_p*. For every* $f \in L^2_p$ *there exist* $\{c_j, j \in \mathcal{N}\}$ *such that*

$$f = \sum_{j=1}^{\infty} c_j u_j \quad \textit{and} \quad c_j = a(f, u_j).$$

Proof. Since T is self-adjoint and compact (Lemma 4.9), we can apply Theorem 3.60 and Theorem 3.61. Moreover, since T is also bounded and positive definite, T has eigenvalues

$$0 \swarrow \ldots \mu_2 \leq \mu_1$$

where we have counted each eigenvalue according to its multiplicity. By Theorem 3.61 the corresponding eigenfunctions

$$u_1, u_2, \ldots$$

can be chosen so that they are orthonormal (with respect to $a(\cdot,\cdot)$) and the span of them is dense in L^2_p. The result then follows from Lemma 3.71. □

4.1.5 Regularity

With the assumption that $\gamma \in PC_p$ (see Definition 3.36), we can derive three results about the regularity of $\mathrm{T}\, u$ when $u \in H^1_p$ and the eigenfunctions of Problem 4.6.

We begin by proving a regularity result for $\mathrm{T}\, u$ when $u \in H^1_p$ that depends on the regularity of $\gamma(\mathbf{x})$. More specifically, we will use Theorem 3.40 which states $\gamma(\mathbf{x}) \in H^{1/2-\epsilon}_p$ for any $\epsilon > 0$ to prove that $\mathrm{T}\, u \in H^{5/2-\epsilon}_p$. Therefore, it is the regularity of $\gamma(\mathbf{x})$ that limits the regularity of $\mathrm{T}\, u$. As well as using Theorem 3.40 to prove the result, we will also use the regularity theory for elliptic boundary value problems that we quoted in Chapter 3. In particular, we use Theorem 3.77 which states that for an elliptic boundary value problem of the form $Lu = f$ on $\mathbb{R}^2$ such that u is periodic (and L has smooth coefficients), if $f \in H^s_p$ for $s \geq 0$, then $u \in H^{s+2}_p$. At first glance it may not seem possible that we can apply this theorem because $\gamma(\mathbf{x})$ is not smooth. However, we will incorporate $\gamma(\mathbf{x})$ into f, leaving L with constant coefficients. This result (Theorem 4.11) is the most important result of this section and our error bounds later in this chapter will rely on it.

The second result is a simple corollary to the first result and is specific for eigenfunctions of Problem 4.6.

The third result is also specific to the eigenfunctions of Problem 4.6. In it we prove that the eigenfunctions of Problem 4.6 are infinitely smooth away from the discontinuities of $\gamma(\mathbf{x})$. Therefore, any limitations on the regularity of the eigenfunctions must come from the behaviour of the eigenfunctions near or at the interface regions. The proof of the third result will use standard regularity theory for elliptic boundary value problems which can be found in [21].

The second and third results about eigenfunctions of Problem 4.6 will allow us to identify an eigenpair of Problem 4.6 with an eigenpair of $\mathrm{L}_{\boldsymbol{\xi}}$ as well as letting us have more insight into the behaviour of the eigenfunctions, even though the results are not required in the rest of this thesis.

Recall our definition of the notation $\lesssim$ from Section 3.1.

Theorem 4.11. *Assume $\gamma \in PC_p$, $u \in H^1_p$ and $\epsilon > 0$. Then $\mathrm{T}\, u \in H^{5/2-\epsilon}_p$ and*

$$\|\mathrm{T}\, u\|_{H^{5/2-\epsilon}_p} \lesssim \|u\|_{H^1_p}$$

where T *is the solution operator corresponding to Problem* **??** *defined the sense of Definition 3.70.*

Proof. Since $\gamma \in PC_p$ (see Definition 3.36) we can use Theorem 3.40 to get $\gamma \in H_p^{1/2-\epsilon'}$ for any $\epsilon' > 0$.

By the definition of T (see Definition 3.70) we have that $w = \mathrm{T}\, u$ is the weak solution of an elliptic boundary value problem of the form

$$\begin{aligned} &Lw = f \qquad \text{on } \mathbb{R}^2 \\ &w \text{ periodic with period cell } \Omega \end{aligned} \tag{4.4}$$

where $L := -(\nabla + i\boldsymbol{\xi})^2 + \mathrm{K}$ and $f := u + \gamma(\mathbf{x})\,\mathrm{T}\, u$. L is an elliptic operator with constant coefficients. Note that we have shifted the term $\gamma(\mathbf{x})\,\mathrm{T}\, u$ onto the right-hand-side of (4.4) so that L has constant coefficients.

The key to completing the proof is to show that $f \in H_p^{1/2-\epsilon}$ and $\|f\|_{H_p^{1/2-\epsilon}} \lesssim \|u\|_{H_p^1}$ so that we can apply Theorem 3.77 to (4.4) to get

$$\|\,\mathrm{T}\, u\|_{H_p^{5/2-\epsilon}} \lesssim \|f\|_{H_p^{1/2-\epsilon}} \lesssim \|u\|_{H_p^1}.$$

By Theorem 3.28 and the definition of f we get

$$\|f\|_{H_p^{1/2-\epsilon}} \lesssim \|u\|_{H_p^{1/2-\epsilon}} + \|\gamma\|_{H_p^{1/2-\epsilon}} \|\,\mathrm{T}\, u\|_{H_p^t} \tag{4.5}$$

for any $t > 1$. We will show that $\mathrm{T}\, u \in H_p^2$. We do this by showing that $f \in L_p^2$ and then use Theorem 3.77 applied to (4.4) to get $\mathrm{T}\, u \in H_p^2$. Since $u \in H_p^1 \subset L_p^2$, $\gamma \in L_p^\infty \subset PC_p$, $\mathrm{T}\, u \in H_p^1 \subset L_p^2$ by definition and T is bounded on H_p^1, it follows that

$$\|f\|_{L_p^2} \lesssim \|u\|_{L_p^2} + \|\gamma\|_\infty \|\,\mathrm{T}\, u\|_{L_p^2} \lesssim \|u\|_{H_p^1} < \infty.$$

Therefore, $f \in L_p^2$, and by Theorem 3.77 applied to (4.4) we get $\mathrm{T}\, u \in H_p^2$ with

$$\|\,\mathrm{T}\, u\|_{H_p^2} \lesssim \|u\|_{H_p^1}.$$

Combining this with (4.5) we get $\|f\|_{H_p^{1/2-\epsilon}} \lesssim \|u\|_{H_p^1}$ and the result follows by applying Theorem 3.77 to (4.4).

In 1D the proof does not require two applications of Theorem 3.77 because the 1D result from Theorem 3.28 for estimating $\|\gamma\,\mathrm{T}\, u\|_{H_p^{1/2-\epsilon}}$ is easier to work with and we can show $\|f\|_{H_p^{1/2-\epsilon}} \lesssim \|u\|_{H_p^1}$ directly. □

Corollary 4.12. *Let* (λ, u) *be an eigenpair of Problem 4.6 with* $\gamma \in PC_p$*. Then for* $\epsilon > 0$ *we get* $u \in H_p^{5/2-\epsilon}$ *and*

$$\|u\|_{H_p^{5/2-\epsilon}} \lesssim \|u\|_{H_p^1}$$

Proof. The result follows directly from Theorem 4.11 using Lemma 3.71 and $\mathrm{T}\,u = \frac{1}{\lambda}u$. □

The following result, although not required in the rest of this chapter, gives us a useful insight into the limitations on the regularity of the eigenfunctions of Problem 4.6.

Theorem 4.13. *With $\gamma \in PC_p$, divide Ω into regions Ω_j, $j = 1, ..., J$ where $\gamma(\mathbf{x})$ is constant. Let (λ, u) be an eigenpair of Problem 4.6. Then*

$$u \in C^\infty(\Omega_j) \qquad \text{for each } j = 1, ..., J.$$

Proof. Let $j \in \{1, ..., J\}$ and let (λ, u) be an eigenpair of Problem 4.6. In each Ω_j we can rewrite Problem 4.6 as an elliptic boundary value problem of the form $Lw = 0$ on Ω_j where $L = \mathrm{L}_{\boldsymbol{\xi}} - \lambda$. L has constant coefficients since $\gamma(\mathbf{x})$ is constant in each Ω_j. $w = u|_{\Omega_j}$ is a weak solution to this boundary value problem and by the definition of Problem 4.6 we have $u \in H^1_p$. Theorem 3 on page 316 of [21] then states that $u \in C^\infty(\Omega_j)$. □

Theorem 4.13 does not include any information about the behaiviour of u on the boundary of each Ω_j, but it does show that if an eigenfunction has a singularity in one of its derivatives, then it must be confined to the *interfaces* of $\gamma(\mathbf{x})$ and it can not "propagate" into regions where $\gamma(\mathbf{x})$ is constant.

4.1.6 Special Case: 1D TE Mode Problem

In this subsection we consider the 1D TE Mode Problem defined by (2.20). We can also think of this problem as being the 1D version of the Scalar 2D Problem that we have been looking at so far in this chapter. In fact, all of the results that we have presented from the Scalar 2D Problem also apply to this 1D problem. We introduce the 1D problem because it is a physically relevant problem in its own right as well as to point out a few results that only hold in 1D or that we were only able to prove in 1D.

Formally, the 1D TE Mode Problem is

$$\frac{d^2h}{dx^2} + \gamma(x)h = \beta^2 h \tag{4.6}$$

where h is the x-component of the magnetic field and β is the component of the wave vector in the z-direction. The coefficient function $\gamma \in PC_p$ is piecewise constant and periodic with period cell $\Omega = [-\frac{1}{2}, \frac{1}{2}]$. We also assume that $0 < \gamma(x) \leq \gamma_{\max}$. We are again interested in finding the eigenfunctions h and the correponding eigenvalues β^2 in (4.6).

We state the problem mathematically as trying to find the spectrum of an operator on a Hilbert space. In this case the Hilbert space is $L^2(\mathbb{R})$ with the usual inner product and the operator is

$$\mathrm{L} = -\frac{d}{dx} - \gamma(x) + \mathrm{K}$$

with domain $H^2(\mathbb{R})$. To obtain L from (4.6) we have multiplied (4.6) by -1 and added a constant K to shift the spectrum into $(0, \infty)$ and ensure that L is postive definite. By the same reasoning as in Theorem 4.1 we have $\sigma(\mathrm{L}) = \sigma_{ess}(\mathrm{L}) \subset \mathbb{R}$. We apply the Floquet Transform to obtain a family of problems: for $\xi \in B := [-\pi, \pi]$ we want to find $\sigma(\mathrm{L}_\xi)$ where

$$\mathrm{L}_\xi := -\left(\frac{d}{dx} + i\xi\right)^2 - \gamma(x) + \mathrm{K}$$

has domain H^2_p and we are now working in the Hilbert space L^2_p. Lemma 4.2 applies to the 1D problem except there is an extension to Part 3 which can be found in Theorem XIII.89 on pages 293 and 294 of [69]. The extension is stated in the following lemma.

Lemma 4.14. *If γ is even then $\lambda(\xi) \in \sigma(\mathrm{L}_\xi)$ considered as a function of ξ is also an even function. Moreover, $\lambda(\xi)$ is continuous and monotone on $[-\pi, 0]$ and $[0, \pi]$.*

This result is a confirmation of Conjecture 4.5 for the 1D case. Since $\lambda(\xi)$ is continuous, even and monotone between 0 and π we can conclude that $\lambda(\xi) \in [\lambda(0), \lambda(\pi)]$ if $\lambda(0) \leq \lambda(\pi)$ and $\lambda(\xi) \in [\lambda(\pi), \lambda(0)]$ if $\lambda(0) > \lambda(\pi)$. Therefore, it is sufficient to only calculate $\sigma(\mathrm{L}_0)$ and $\sigma(\mathrm{L}_\pi)$ to determine $\sigma(\mathrm{L})$ (see Theorem 3.63).

We are now free to concentrate on calculating $\sigma(\mathrm{L}_\xi)$ for a fixed $\xi \in B$. We write down the variational problem corresponding to finding an eigenvalue of $\sigma(\mathrm{L}_\xi)$ and corresponding eigenfunction.

Problem 4.15. For a fixed $\xi \in B$, find an eigenpair (λ, u) where $\lambda \in \mathbb{C}$ and $0 \neq u \in H^1_p$ such that

$$a(u, v) = \lambda b(u, v) \qquad \forall v \in H^1_p \tag{4.7}$$

where

$$\begin{aligned} a(u, v) &= \int_\Omega \left(\tfrac{d}{dx} + i\xi\right) u \overline{\left(\tfrac{d}{dx} + i\xi\right) v} + (\mathrm{K} - \gamma)\, u\overline{v} dx \\ b(u, v) &= \int_\Omega u\overline{v} dx. \end{aligned}$$

This variational problem is just the 1D version of Problem 4.6. We can prove that $a(\cdot, \cdot)$ is bounded, coercive and Hermitian in the same way as in Lemma 4.7 and it follows that $a(\cdot, \cdot)$ defines an inner product on H^1_p with $\|\cdot\|_a := a(\cdot, \cdot)^{1/2}$ defining the induced norm. We can also define a solution operator T for the 1D problem. It has the same properties as T for the 2D problem and we can deduce the same properties of the spectrum of Problem 4.15 as we could for the spectrum of Problem 4.6.

We can also follow the same proof as in Subsection 4.1.5 to show that the eigenfunctions of Problem 4.15 have $H_p^{5/2-\epsilon}$ regularity for every $\epsilon > 0$. However, we can also prove a slightly different regularity result for the eigenfunctions of Problem 4.15. We get the following Theorem.

Theorem 4.16. *Let $u \in H_p^1$. Then $\mathrm{T}\,u$ and $(\mathrm{T}\,u)'$ are absolutely continuous and $(\mathrm{T}\,u)''$ is continuous except where $\gamma(x)$ is discontinuous and is absolutely continuous on the intervals of continuity.*

Proof. As in Theorem 4.11 we define a boundary value problem $Lw = f$ on $\mathbb{R}$ such that w is periodic with period cell Ω and where $L := -(\nabla + i\boldsymbol{\xi})^2 + \mathrm{K}$ and $f := u + \gamma(\mathbf{x})\,\mathrm{T}\,u$. L is an elliptic operator with constant coefficients and $f \in L_p^2$. $w = \mathrm{T}\,u$ is a weak solution to $Lw = f$. Therefore, using Theorem 3.77, $\mathrm{T}\,u \in H_p^2$. This implies that $(\mathrm{T}\,u)'' \in L_p^2$. It then follows that $(\mathrm{T}\,u)'' \in L^1(\Omega)$ since $L_p^2 \subset L^2(\Omega) \subset L^1(\Omega)$. Next, we use Lemma 7.3.5 on page 317 of [4] to get $(\mathrm{T}\,u)'$ is absolutely continuous. It also follows that $\mathrm{T}\,u$ is absolutely continuous. Now we apply integration by parts to

$$a(\mathrm{T}\,u, \phi) = b(u, \phi) \qquad \forall \phi \in C_0^\infty(\Omega)$$

to get

$$\int_\Omega \left((\tfrac{d}{dx} + i\xi)^2\,\mathrm{T}\,u - (\mathrm{K} - \gamma)\,\mathrm{T}\,u + u \right) \overline{\phi} dx = 0 \qquad \forall \phi \in C_0^\infty(\Omega).$$

Therefore, $(\mathrm{T}\,u)'' = -2i\xi(\mathrm{T}\,u)' + (\xi^2 + \mathrm{K} - \gamma(x))\,\mathrm{T}\,u - u$ almost everywhere. It then follows that $(\mathrm{T}\,u)''$ is continuous except at the discontinuities of $\gamma(x)$ and absolutely continuous on the intervals of continuity. □

It follows, just as in Corollary 4.12, that if u is an eigenfunction of Problem 4.15, then u and u' are absolutely continuous and u'' is continuous except where $\gamma(x)$ is discontinuous and is absolutely continuous on the intervals of continuity.

4.1.7 Examples

In this subsection we define 1D and 2D model problems that we will use in numerical computations to verify our theoretical results in the rest of this chapter.

In all of the model problems $\gamma(\mathbf{x})$ will have two possible values, $\gamma_a = 157.9$ or $\gamma_g = 309.5$. These two values of γ correspond to a photonic crystal fibre that is made from glass and air with refractive indices of 1.4 and 1 respectively. In all of the model problems we have fixed the period cell of the cladding structure so that it has a period cell of length 1, and we are considering light that has a wavelength that is half of the cladding period cell width, i.e. $\lambda_0 = \frac{1}{2}$, for all of the model problems. Also, in all of our model problems we have chosen $\gamma(\mathbf{x})$ to be an even function. This is because real

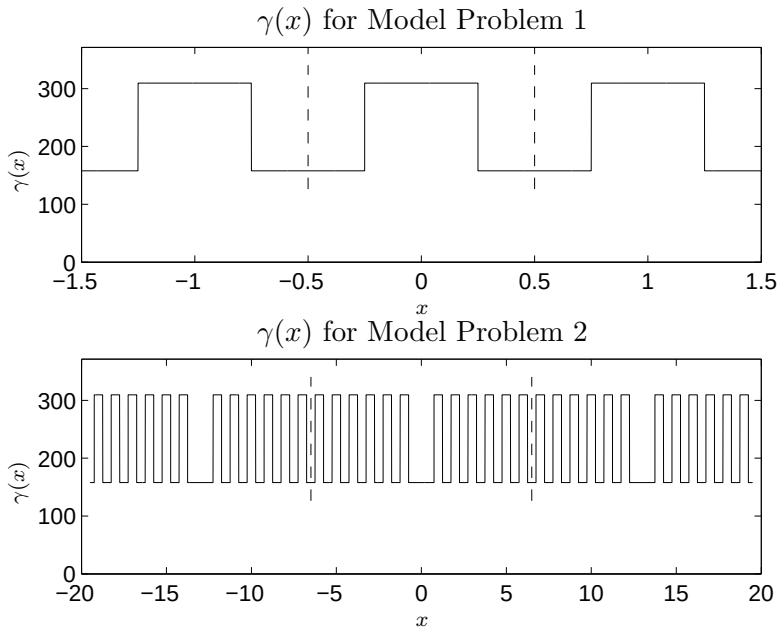

Figure 4-1: Plot of $\gamma(x)$ for Model Problems 1 and 2. Notice that the period cell of $\gamma(x)$ in Model Problem 1 is the same length as a cell in the cladding of Model Problem 2.

PCFs usually have some form of symmetry and since all of our PCFs have a square structure even symmetry is the natural choice of symmetry.

Model Problem 1 is a 1D problem where $\gamma(x)$ is describing a pure photonic crystal that has a 50:50 glass to air ratio and a period cell $\Omega = (-1/2, 1/2)$. Figure 4-1 has a plot of $\gamma(x)$ for Model Problem 1.

Model Problem 2 models a 1D PCF by using the supercell method. $\gamma(x)$ describes the cladding structure together with a central defect where there are 12 period cells of cladding between each defect. For this problem $\Omega = (-\frac{13}{2}, \frac{13}{2})$ and $B = [-\frac{\pi}{13}, \frac{\pi}{13}]$. The reason $\Omega \neq (-\frac{1}{2}, \frac{1}{2})$ is so that if we removed the defect in the supercell of $\gamma(x)$ for Model Problem 2 then $\gamma(x)$ would be exactly the same as in Model Problem 1. Put another way, a cell in the cladding of $\gamma(x)$ of Model Problem 2 is exactly the same as a period cell of $\gamma(x)$ from Model Problem 1. This will ensure that the band gaps in Model Problem 1 are the same as the band gaps in Model Problem 2. A theoretical justification for the band gaps remaining unchanged is given in Part 4 of Theorem 3.60. Figure 4-1 has a plot of $\gamma(x)$ for Model Problem 2.

Model Problem 3 is a 2D version of Model Problem 1. Again, $\gamma(\mathbf{x})$ describes a photonic crystal. It consists of glass with square air holes. Figure 4-2 has a diagram of the period cell for $\gamma(\mathbf{x})$ in this problem.

Model Problem 4 is a 2D version of Model Problem 2 except that the cladding in

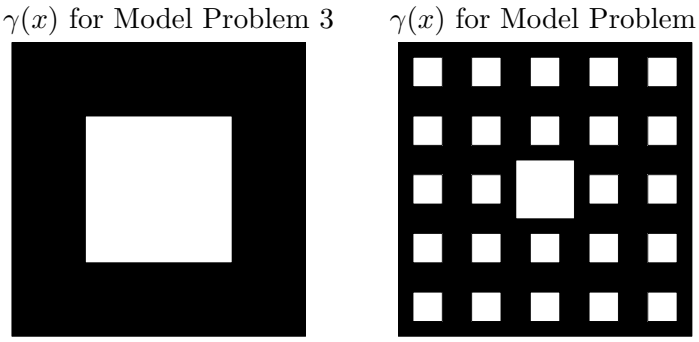

Figure 4-2: Plot of $\gamma(\mathbf{x})$ for Model Problems 3 and 4. The scale of $\gamma(\mathbf{x})$ in Model Problem 4 is such that a period cell from Problem 3 is the same length as a cell in the cladding of $\gamma(\mathbf{x})$ in Problem 4. The black regions are glass and the white regions correspond to air holes.

Model Problem 4 has fewer cells. $\gamma(\mathbf{x})$ has a 5x5 supercell with a central defect. The reason we have chosen a supercell with fewer cells between each defect than Model Problem 2 is to make this problem easier to solve. $\gamma(\mathbf{x})$ represents a PCF in this problem and Figure 4-2 has a diagram of the period cell of $\gamma(\mathbf{x})$ for this problem.

Since Problems 1 and 3 correspond to pure photonic crystal we want to accurately calculate the band gaps for these problems (see Chapter 2 for a discussion of the background physics). Therefore, we will be interested in the convergence of our numerical method for all of the eigenvalues that lie in the interval $[0, \gamma_g]$. For Problem 1 this requires the first 5 eigenvalues whereas Problem 3 requires the first 22 eigenvalues. The bands for Problems 1 and 3 are plotted in Figures 4-3 and 4-4. The bands are constructed by solving the Floquet transformed problem for a range of $\boldsymbol{\xi} \in B$. This idea is represented by plotting the eigenvalues of the Floquet transformed problem against $\boldsymbol{\xi}$. The lines are then projected onto the vertical axis to construct the bands. For Problem 1 in Figure 4-3 we have taken $\xi \in B = [-\pi, \pi]$, although the plot confirms Lemma 4.14, that we only need to do calculations for $\xi = 0$ and $\xi = \pi$. For Problem 3 we take $\boldsymbol{\xi} \in \partial B_I$ where B_I is an irreducible Brillouin zone to construct the bands ($\gamma(\mathbf{x})$ has horizontal, vertical and diagonal mirror symmetry). For Problem 3, the boundary of the irreducible Brillouin zone ∂B_I is the boundary of a triangle with vertices $(0, 0)$, $(0, \frac{\pi}{13})$ and $(\frac{\pi}{13}, \frac{\pi}{13})$. In this thesis we are interested in the convergence of our numerical method and we will take $\boldsymbol{\xi} = (0, 0)$ and $\boldsymbol{\xi} = (\pi, \pi)$ as representative examples for the rest of our computations (except in Figure 4-4).

Model Problems 2 and 4 are supercell problems and they are attempting to model a PCF with a central defect that is surrounded by photonic crystal. The cladding for

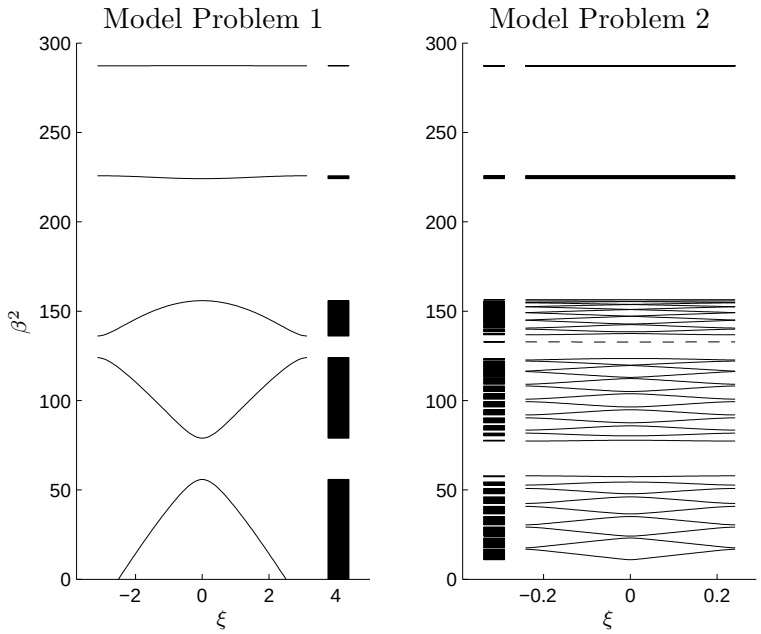

Figure 4-3: A plot of the spectra of Model Problems 1 and 2. The spectra are represented with solid black blocks (or bands) running vertically nearest the middle of the page. Each band is constructed by projecting the corresponding line onto the vertical axis. And each line is an eigenvalue of the Floquet transformed problem as a function of $\xi \in B$, i.e. $\lambda(\xi)$. Problem 1 has five bands in the interval $[0, \gamma_g]$. Problem 2 has approximately the same band gaps as Model Problem 2 except there appears to be an isolated eigenvalue (38th from top) in the third band gap (dashed line). For each band in Problem 1 there are approximately 13 bands in Problem 2. This corresponds to the number of cells in the supercell of Problem 2. There are small band gaps between every band of Problem 2 but these small gaps arise from having a supercell with finite cladding.

Problem 2 is the photonic crystal in Problem 1 and the cladding for Problem 4 is the photonic crystal in Problem 3. By this we mean that a period cell of $\gamma(x)$ in Problem 1 is the same as a cell of the cladding in Problem 2. Likewise for Problems 3 and 4. We expect the bands of Problems 2 and 4 to approximate the bands of Problems 1 and 3 respectively (see Figure 4-3). Indeed, if we changed Problems 2 and 4 so that there is more cladding between the defects in the structure of $\gamma(\mathbf{x})$ then the bands of Problem 2 and 4 would provide a better approximation of the bands of Problems 1 and 3 (see discussion of supercell method in Chapter 2). Therefore, once we have located the band gaps for Problems 1 and 3 we will search for guided modes of Problems 2 and 4 that lie in these band gaps. We can see in Figure 4-3 that in Problem 2 the 38th eigenvalue appears to be an isolated eigenvalue. In Figure 4-4 we can see that there is a band gap in the interval $[279.6259, 286.9147]$ and this is where we will search for

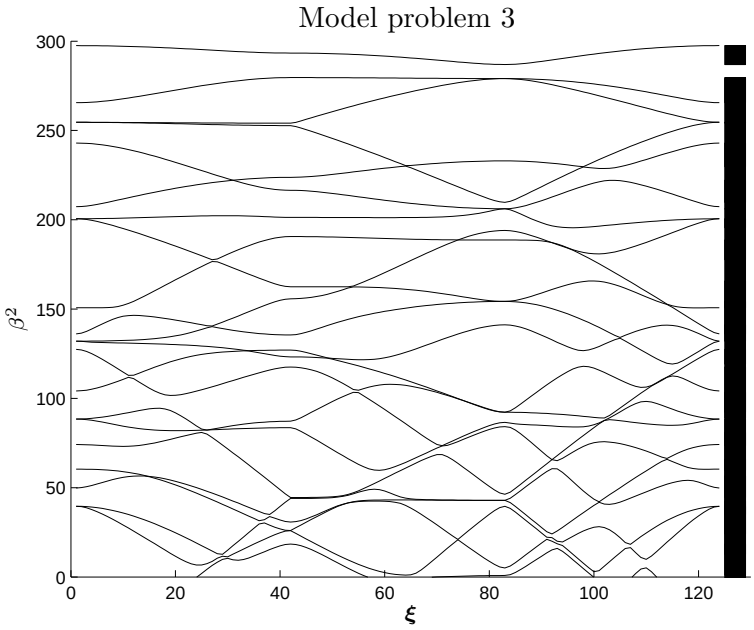

Figure 4-4: A plot of the spectrum of model problem 3. The spectrum is represented with the solid black vertical bands on the right. These bands are the projection of all of the lines onto the vertical axis. Model problem 3 only has one band gap, the interval $[279.6259, 286.9147]$. The horizontal axis of the plot is a parameterization of $\boldsymbol{\xi}$ as it runs around the edge of B_I, a triangle with vertices $(0,0)$, $(\frac{\pi}{13},0)$ and $(\frac{\pi}{13},\frac{\pi}{13})$.

guided modes in Problem 4. Since the band gap in Problem 3 is after the first band we expect the possible guided mode to be approximately the 25th eigenvalue in Problem 4.

The usual technique for searching for a guided mode in a band gap is to use a "shift-invert" strategy to find the the eigenvalue closest to the middle of the gap. However, since the number of eigenvalues up to the guided mode is not too large for these problems this is not the only strategy available to us. Alternatively, we can compute all of the eigenvalues up to and including the possible guided mode. This is the strategy that we will use since in the next section we find that the matrix from the discretization method is positive definite and we can use PCG instead of GMRES to solve linear systems in the implementation when the "shift-invert" strategy is not used. We will calculate the first 30 eigenvalues of Model Problem 4.

4.2 Standard Spectral Galerkin Method

In this section we describe the basic method that we have chosen to use and analyze for approximating the spectrum of $\mathrm{L}_{\boldsymbol{\xi}}$ for a fixed $\boldsymbol{\xi} \in B$. It is a spectral Galerkin

method, but it is more commonly referred to as the *plane wave expansion method*. The method replaces the infinite dimensional Problem 4.6 with a finite dimensional problem that we represent as a matrix eigenvalue problem. The matrix eigenvalue problem is solved using existing iterative techniques. As well as presenting details for the efficient implementation, the main focus is the error analysis for the method. We also support our theory with numerical examples.

The section is divided into four subsections. In the first subsection we describe the method. In the second subsection we give some details relating to the efficient implementation of the method as well as defining a preconditioner matrix and proving a result about our preconditioner. In the third subsection we present our main error bounds and in the fourth subsection we present the results from some numerical computations for our model problems.

4.2.1 The Method

In this subsection we apply a spectral Galerkin method to Problem 4.6 to get a finite dimensional problem.

For $G \in \mathbb{N}$ we choose a finite dimensional space $S_G \subset H^1_p$ and apply the Galerkin method (see Definition 3.72) to Problem 4.6. We refer to this method as a *spectral* Galerkin method because we construct S_G from functions that have global support in Ω. The method is not a spectral method in the sense that the finite dimensional space consists of functions that are eigenfunctions of $\mathrm{L}_{\boldsymbol{\xi}}$. More specifically, we define

$$S_G := S_G^{(2)} = \operatorname{span}\{\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} : \mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}\} \tag{4.8}$$

where $\mathbb{Z}^2_{G,\mathrm{o}} = \{\mathbf{n} \in \mathbb{Z}^2 : |\mathbf{n}| \leq G\}$ (see Subsection 3.2.3). We also denote the dimension of S_G by $N := \dim \mathcal{S}_G = \mathcal{O}(G^2)$. Applying the Galerkin method to Problem 4.6 gives us the following discrete variational eigenvalue problem

Problem 4.17. Find $\lambda_G \in \mathbb{R}$ and $0 \neq u_G \in \mathcal{S}_G$ such that

$$a(u_G, v_G) = \lambda_G b(u_G, v_G) \qquad \forall v_G \in \mathcal{S}_G. \tag{4.9}$$

This problem, since it is finite dimensional, can be rewritten as a matrix eigenvalue problem. We do this by first expanding u_G in terms of a basis for $\mathcal{S}_G$. This expansion is just the Fourier Series of u_G,

$$u_G(\mathbf{x}) = \sum_{\mathbf{g}\in\mathbb{Z}^2_{G,\mathrm{o}}} u_{\mathbf{g}}\, \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} \tag{4.10}$$

where the coefficients of the expansion are the Fourier coefficients of u_G, $u_{\mathbf{g}} = [u_G]_{\mathbf{g}}$.

Since the functions $\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}$, with $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$ form a basis for $\mathcal{S}_G$, it is sufficient to

only choose $v_G = \mathrm{e}^{i2\pi\mathbf{g}'\cdot\mathbf{x}}$ for $\mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}}$ as test functions in (4.9). Restricting the test functions v_G to this finite number of possiblities, Problem 4.17 is equivalent to

$$\sum_{\mathbf{g}\in\mathbb{Z}^2_{G,\mathrm{o}}} u_{\mathbf{g}} a(\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}, \mathrm{e}^{i2\pi\mathbf{g}'\cdot\mathbf{x}}) = \lambda_G \sum_{\mathbf{g}\in\mathbb{Z}^2_{G,\mathrm{o}}} u_{\mathbf{g}} b(\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}, \mathrm{e}^{i2\pi\mathbf{g}'\cdot\mathbf{x}}) \qquad \forall \mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}}. \tag{4.11}$$

Now define a one-to-one map $i : \mathbb{Z}^2_{G,\mathrm{o}} \to \{n \in \mathbb{N} : n \leq N\}$ that orders $\mathbb{Z}^2_{G,\mathrm{o}}$ in ascending order of magnitude, i.e. $i(\mathbf{g}) < i(\mathbf{g}')$ if $|\mathbf{g}| < |\mathbf{g}'|$. Using this map we can define a vector $\mathbf{u}$ of length N that contains all of the Fourier coefficients in the expansion of u_G in (4.10). The entries of $\mathbf{u}$, are defined as

$$u_{i(\mathbf{g})} = u_{\mathbf{g}} = [u_G]_{\mathbf{g}} \qquad \forall \mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}.$$

Now define a $N \times N$ matrix A with entries defined by

$$\begin{aligned}
\mathrm{A}_{i(\mathbf{g}'),i(\mathbf{g})} &= a(\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}, \mathrm{e}^{i2\pi\mathbf{g}'\cdot\mathbf{x}}) && (4.12)\\
&= \int_\Omega (\nabla + i\boldsymbol{\xi})\, \mathrm{e}^{i2\pi\mathbf{g}'\cdot\mathbf{x}} \cdot \overline{(\nabla + i\boldsymbol{\xi})\, \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}} + (\mathrm{K} - \gamma)\, \mathrm{e}^{i2\pi\mathbf{g}'\cdot\mathbf{x}}\, \overline{\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}} dx \\
&= (i\xi + i2\pi\mathbf{g}') \cdot (-i\xi - i2\pi\mathbf{g}) \int_\Omega \mathrm{e}^{i2\pi(\mathbf{g}'-\mathbf{g})\cdot\mathbf{x}}\, dx \\
&\quad + \mathrm{K} \int_\Omega \mathrm{e}^{i2\pi(\mathbf{g}'-\mathbf{g})\cdot\mathbf{x}}\, dx - \int_\Omega \gamma(\mathbf{x})\, \mathrm{e}^{i2\pi(\mathbf{g}'-\mathbf{g})\cdot\mathbf{x}}\, dx \\
&= \left(|\xi + 2\pi\mathbf{g}|^2 + \mathrm{K}\right) \delta_{i(\mathbf{g}),i(\mathbf{g}')} - [\gamma]_{\mathbf{g}-\mathbf{g}'} \qquad \forall \mathbf{g}, \mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}}. && (4.13)
\end{aligned}$$

If we use this together with the fact that

$$b(\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}, \mathrm{e}^{i2\pi\mathbf{g}'\cdot\mathbf{x}}) = \delta_{i(\mathbf{g}),i(\mathbf{g}')} \qquad \forall \mathbf{g}, \mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}}$$

we can write (4.11) as a matrix eigenvalue problem

$$\mathrm{A}\,\mathbf{u} = \lambda_G \mathbf{u}. \tag{4.14}$$

The matrix A has a special form due to our choice of basis functions of $\mathcal{S}_G$. Since $\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}$ are eigenfunctions of the Laplacian and since they are orthogonal with respect to the $L^2(\Omega)$ inner product we can see in (4.13) that A has a special form. It can be expanded as $\mathrm{A} = \mathrm{D} - \mathrm{V}$ where D is a diagonal matrix with diagonal entries given by $\mathrm{D}_{i(\mathbf{g}),i(\mathbf{g})} = |\boldsymbol{\xi} + 2\pi\mathbf{g}|^2 + \mathrm{K}$ and V is a dense matrix with entries given by $\mathrm{V}_{i(\mathbf{g}),i(\mathbf{g}')} = [\gamma]_{\mathbf{g}-\mathbf{g}'}$. For a given vector $\mathbf{v} \in \mathbb{R}^N$, it is obvious that $\mathrm{D}\,\mathbf{v}$ can be computed very quickly since D is diagonal but it is not immediately obvious how $\mathrm{V}\,\mathbf{v}$ can be computed quickly.

The matrix V contains the Fourier coefficients of $\gamma(\mathbf{x})$ whereas the vector $\mathbf{v}$ contains the Fourier coefficients of another function. In a certain sense, the product $\mathrm{V}\,\mathbf{v}$ represents the multiplication of $\gamma(\mathbf{x})$ and this other function, and this multiplication

can be computed efficiently using the Fast Fourier Transform. This is the topic of the next subsection.

Now we prove that A is Hermitian and positive definite. If $\gamma(\mathbf{x})$ is an even function then the Fourier coefficients of $\gamma(\mathbf{x})$ are real and A will be a real matrix (see (4.13)). Therefore, A Hermitian implies that A is symmetric. All of our model problems from Section 4.1.7 have even $\gamma(\mathbf{x})$ and so we will refer to A as being symmetric positive definite in the rest of this chapter. The proof relies on the fact that $a(\cdot,\cdot)$ is coercive and Hermitian.

Theorem 4.18. *The matrix* A *from* (4.14) *is Hermitian and positive definite.*

Proof. First, we show that A is Hermitian. From (4.12) and $a(\cdot,\cdot)$ Hermitian we get

$$\mathrm{A}_{i(\mathbf{g}),i(\mathbf{g}')} = a(\mathrm{e}^{i2\pi\mathbf{g}'\cdot\mathbf{x}}, \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}) = \overline{a(\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}, \mathrm{e}^{i2\pi\mathbf{g}'\cdot\mathbf{x}})} = \overline{\mathrm{A}_{i(\mathbf{g}'),i(\mathbf{g})}} \qquad \forall \mathbf{g},\mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}}.$$

Therefore, A is Hermitian.

Now we show that A is postive definite. Let $\mathbf{x} \in \mathbb{C}^N$ such that $\mathbf{x} \neq 0$ and define $\mathcal{X} \in \mathcal{S}_G$ by

$$\mathcal{X}(\mathbf{x}) = \sum_{\mathbf{g}\in\mathbb{Z}^2_{G,\mathrm{o}}} x_{i(\mathbf{g})}\, \mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}\,.$$

From (4.12) and $a(\cdot,\cdot)$ coercive we then get

$$\begin{aligned}
\mathbf{x}^H\, \mathrm{A}\, \mathbf{x} &= \sum_{\mathbf{g},\mathbf{g}'\in\mathbb{Z}^2_{G,\mathrm{o}}} \mathrm{A}_{i(\mathbf{g}')i(\mathbf{g})}\, \overline{x_{i(\mathbf{g}')}} x_{i(\mathbf{g})} \\
&= \sum_{\mathbf{g},\mathbf{g}'\in\mathbb{Z}^2_{G,\mathrm{o}}} a(\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}}, \mathrm{e}^{i2\pi\mathbf{g}'\cdot\mathbf{x}})\overline{x_{i(\mathbf{g}')}} x_{i(\mathbf{g})} \\
&= a(\mathcal{X},\mathcal{X}) \gtrsim \|\mathcal{X}\|_{H^1_p} > 0.
\end{aligned}$$

□

Before we move onto the implementation of our method let us discuss the 1D problem and the matrix eigenproblem that is derived in that case.

For the 1D problem we define $\mathcal{S}_G := \mathcal{S}^{(1)}_G$ as in Subsection 3.2.3. We apply the Galerkin method with $\mathcal{S}_G$ replacing H^1_p to obtain a discrete variational problem as in Problem 4.17. We then write down a $N \times N$ matrix eigenvalue problem that is equivalent to the discrete variational problem where $N = 2G + 1$. The only difference from the 2D formulation is that instead of using $i(\cdot)$ to define an ordering for the matrix and vector entries we order the matrix and vector entries from $-G$ to G. For example, $\mathbf{u}$ is now a N vector

$$\mathbf{u} = [u_{-G} \ldots u_{-1}\; u_0\; u_1 \ldots u_G]^T, \tag{4.15}$$

the diagonal entries of D are given by $\mathrm{D}_{ii} = (\xi^2 + 2\pi(i - G - 1))^2 + \mathrm{K}$ and the entries of V

are given by $\mathrm{V}_{ij} = [\gamma]_{i-j}$, for $i, j = 1, ..., 2G+1$. We see that V is a Toeplitz matrix and we know from [84] (Algorithm 4.2.2 on page 209) that Toeplitz matrix vector products may be computed in $\mathcal{O}(N \log N)$ operations using the Fast Fourier Transform as in the 2D case.

We return to discussing the 2D problem in the next subsection.

4.2.2 Implementation

In this subsection we discuss our method for solving the matrix eigenvalue problem (4.14). Again, the general discussion will be for the 2D problem with particular comments about the 1D problem where necessary. We frequently refer to theory that was presented in Section 3.6.

We want to find the eigenvalues of A (from (4.14)) in the interval $[0, \mathrm{K}]$ and corresponding eigenfunctions. Since A is a positive definite matrix (Theorem 4.18), this corresponds to the smallest eigenvalues of A up to K. We use a Krylov subspace iterative method since we are not interested in computing all of the eigenvalues of A. Indeed, it would be too costly to compute all of them when N is large. More specifically, we use the Implicitly Restarted Arnoldi's (IRA) method applied to A^{-1}.

The IRA method applied to A was our first choice for calculating the smallest eigenvalues of A because it approximates the extremal eigenvalues of a matrix. However, the matrix A has many well-spaced, very large eigenvalues and the smallest eigenvalues of A are clustered. This causes the IRA method applied to A to approximate the largest eigenvalues of A better than the smallest eigenvalues of A. Applying the IRA method to A^{-1} reverses this situation.

At each step or iteration of the IRA method we require the operation of A^{-1}. This is obtained by solving a linear system with coefficient matrix A. Since A is symmetric and positive definite (spd) (Theorem 4.18) we can use the preconditioned conjugate gradient method (PCG). PCG only requires scalar-vector multiplication, vector-vector addition and matrix-vector multiplications. Of these three operations, matrix-vector multiplications are potentially the most costly as scalar-vector multiplication and vector-vector addition only require $\mathcal{O}(N)$ operations. We improve the performance of PCG by using a preconditioner that is effective at limiting the number of iterations required in PCG to $\mathcal{O}(1)$ as well as using an algorithm that can compute matrix-vector products in $\mathcal{O}(N \log N)$ operations. All together, we obtain the operation of A^{-1} in $\mathcal{O}(N \log N)$ operations. This is a big improvement over a direct method such as Gauss elimination which would require $\mathcal{O}(N^3)$ operations to solve a system with A. Our method also improves on the amount of storage required to compute the operations of A^{-1}. Gauss elimination requires the storage of every non-zero entry of A. For our problem this would be N^2 entries since A is dense. Our algoritm only requires $\mathcal{O}(N)$ entries to store A since $\mathrm{A} = \mathrm{D} - \mathrm{V}$ where D is a diagonal matrix and V is a matrix with only $\mathcal{O}(N)$

distinct entries.

In this subsection we present the algorithm that can compute matrix-vector products with A in $\mathcal{O}(N \log N)$ operations, define a preconditioner for A and prove a result that shows the optimality of the preconditioner. We begin with the algorithm for computing matrix-vector products.

Since $\mathrm{A} = \mathrm{D} - \mathrm{V}$ where D is diagonal, and matrix-vector products with diagonal matrices can be computed in $\mathcal{O}(N)$ operations, we need a fast algorithm for matrix-vector products with V. The algorithm presented below uses the Fast Fourier Transform (FFT) to compute the matrix vector product with V for the 2D problem. It is essentially an algorithm for computing the convolution of two Fourier Series.

In this section N_f defines is the size of the space that the FFT operates on and in the algorithm below we must choose $N_f \geq 4G + 1$. To get the best performance from the FFT we want to choose $N_f = 2^n$ for some $n \in \mathbb{N}$. In practice we fix N_f, and then we choose $G = N_f/4 - 1$. N is then determined by the number of elements in $\mathbb{Z}^2_{G,\mathrm{o}}$. Note that N represents the number of degrees of freedom in the discrete problem and is $\mathcal{O}(G^2)$ for the 2D problem which we are currently discussing.

We now make a remark about the notation used in the algorithm that follows. Capital letters $X, Y, \widehat{X}, \widehat{Y}$ are all $N_f \times N_f$ matrices that represent functions in $\mathcal{T}^{(2)}_{N_f}$. X, Y store nodal values of functions in $\mathcal{T}^{(2)}_{N_f}$ while $\widehat{X}, \widehat{Y}$ store Fourier coefficients of functions in $\mathcal{T}^{(2)}_{N_f}$. The indexing convention is the same as in Subsection 3.2.4, i.e. for $f \in \mathcal{T}^{(2)}_{N_f}$ we write

$$X_{ij} = f(\tfrac{1}{N_f}((i,j) - \mathbf{g}_0))$$
$$\widehat{X}_{ij} = [f]_{(i,j)-\mathbf{g}_0}$$

for all $i, j = 1, \ldots, N_f$ where $\mathbf{g}_0 := (\frac{N_f}{2} + 1, \frac{N_f}{2} + 1) = (2G + 3, 2G + 3)$.

We also let $\mathrm{fft}(\cdot)$ and $\mathrm{ifft}(\cdot)$ denote the 2D FFT and the 2D Inverse FFT respectively, as in Subsection 3.2.4, so that $\widehat{X} = \mathrm{fft}(X)$ and $X = \mathrm{ifft}(\widehat{X})$.

Algorithm 4.19. Let $\mathbf{x}$ be a vector of length N and let $\widehat{Y}$ be the $N_f \times N_f$ matrix of Fourier coefficients of γ such that $\widehat{Y}_{ij} = [\gamma]_{(i,j)-\mathbf{g}_0}$ for $i, j = 1, \ldots, N_f$. Pre-compute $Y \leftarrow \mathrm{ifft}(\widehat{Y})$. The following algorithm computes a new vector that is denoted, $\mathcal{V}(\mathbf{x})$.
$\widehat{X}_{ij} \leftarrow 0$ for $i, j = 1, \ldots, N_f$
$\widehat{X}_{\mathbf{g}+\mathbf{g}_0} \leftarrow \mathbf{x}_{i(\mathbf{g})}$ for every $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$
$X \leftarrow \mathrm{ifft}(\widehat{X})$
$X_{ij} \leftarrow Y_{ij} X_{ij}$ for $i, j = 1, \ldots, N_f$
$\widehat{X} \leftarrow \mathrm{fft}(X)$
$(\mathcal{V}(\mathbf{x}))_{i(\mathbf{g})} \leftarrow \widehat{X}_{\mathbf{g}+\mathbf{g}_0}$ for every $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.

The main cost of this algorithm are the Fast Fourier Transforms which are com-

puted in $\mathcal{O}(N_f^2 \log N_f)$ operations ($\mathcal{O}(N \log N)$ since $N = \mathcal{O}(N_f^2)$). In practice, each application of the algorithm uses one inverse FFT and one FFT. The inverse FFT $Y \leftarrow \text{ifft}(\widehat{Y})$ is usually computed only once in the setup and then stored for use when the algorithm is applied repeatedly.

We can view Algorithm 4.19 as an algorithm that converts the Fourier coefficients in $\mathbf{x}$ and V into real space; multiplies the two functions together in real space; then converts the real space data back into Fourier space; before finally, discarding unwanted high frequency components. We will use results from Subsection 3.2.5 and [72] to prove that the action of Algorithm 4.19 is equal to matrix-vector multiplication by the matrix V.

Theorem 4.20. $\mathcal{V}(\mathbf{x}) = \mathrm{V}\,\mathbf{x}$ *for all* $\mathbf{x} \in \mathbb{C}^N$.

Proof. Recall from Subsection 3.2.3 that

$$\mathbb{Z}^2_{G,\mathrm{o}} = \left\{\mathbf{n} \in \mathbb{Z}^2 : |\mathbf{n}| \leq G\right\}$$
$$\mathbb{Z}^2_{G,\square} = \left\{\mathbf{n} \in \mathbb{Z}^2 : -\tfrac{G}{2} \leq n_i < \tfrac{G}{2}, i = 1, 2\right\}.$$

Let $\mathbf{x} \in \mathbb{C}^N$ and define $\mathcal{X} \in \mathcal{S}^{(2)}_G$ by

$$\mathcal{X}(\mathbf{t}) := \sum_{\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}} \mathbf{x}_{i(\mathbf{g})}\, \mathrm{e}^{i2\pi \mathbf{g}\cdot\mathbf{t}} \qquad \forall \mathbf{t} \in \mathbb{R}^2.$$

We will also define

$$\mathcal{Y}(\mathbf{t}) := \mathrm{P}^{(\mathcal{T})}_{N_f}\, \gamma(\mathbf{t}) = \sum_{\mathbf{g} \in \mathbb{Z}^2_{G,\square}} [\gamma]_{\mathbf{g}}\, \mathrm{e}^{i2\pi \mathbf{g}\cdot\mathbf{t}} \qquad \forall \mathbf{t} \in \mathbb{R}^2$$

where $\mathrm{P}^{(\mathcal{T})}_{N_f}$ is the projection onto $\mathcal{T}^{(2)}_{N_f}$ defined in Subsection 3.2.5. Recall that $[\cdot]_{\mathbf{g}}$ denotes the Fourier coefficient with index $\mathbf{g}$ and let $(\cdot)_n$ denote the n-th entry of a vector. We also use the projection onto $\mathcal{T}^{(2)}_{N_f}$ that is based on the nodal values of a function, Q_{N_f}. This projection is also defined in Subsection 3.2.5.

The proof is divided into three parts:

1. $(\mathrm{V}\,\mathbf{x})_{i(\mathbf{g})} = [\mathcal{X}\mathcal{Y}]_{\mathbf{g}}$ for all $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.

2. $[\mathcal{X}\mathcal{Y}]_{\mathbf{g}} = [\mathrm{Q}_{N_f}(\mathcal{X}\mathcal{Y})]_{\mathbf{g}}$ for all $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.

3. $[\mathrm{Q}_{N_f}(\mathcal{X}\mathcal{Y})]_{\mathbf{g}} = (\mathcal{V}(\mathbf{x}))_{i(\mathbf{g})}$ for all $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.

Part 1. For $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$,

$$\begin{aligned}
(\mathrm{V}\,\mathbf{x})_{g(\mathbf{g})} &= \sum_{\mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}}} \mathrm{V}_{g(\mathbf{g})g(\mathbf{g}')}\, \mathbf{x}_{g(\mathbf{g}')} \\
&= \sum_{\mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}}} [\gamma]_{\mathbf{g}-\mathbf{g}'} [\mathcal{X}]_{\mathbf{g}'} && \text{by definition of V} \\
&= \sum_{\mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}}} [\mathcal{Y}]_{\mathbf{g}-\mathbf{g}'} [\mathcal{X}]_{\mathbf{g}'} && \text{by definition of } \mathcal{Y} \\
&= \sum_{\mathbf{g}' \in \mathbb{Z}^2} [\mathcal{Y}]_{\mathbf{g}-\mathbf{g}'} [\mathcal{X}]_{\mathbf{g}'} && \text{since } \mathcal{X} \in \mathcal{S}^{(2)}_G \\
&= [\mathcal{X}\mathcal{Y}]_{\mathbf{g}} && \text{by Theorem 28 on page 23 of [36].}
\end{aligned}$$

Part 2. According to Lemma 3.31 we have,

$$[\mathrm{Q}_{N_f}(\mathcal{X}\mathcal{Y})]_{\mathbf{g}} = \sum_{\mathbf{g}' \in \mathbb{Z}^2} [\mathcal{X}\mathcal{Y}]_{\mathbf{g}+N_f\mathbf{g}'} \qquad \text{for } \mathbf{g} \in \mathbb{Z}^2_{N_f,\square}. \tag{4.16}$$

Now observe that since $\mathcal{X} \in \mathcal{S}^{(2)}_G \subset \mathcal{T}^{(2)}_{2G}$ and $\mathcal{Y} \in \mathcal{T}^{(2)}_{N_f}$, we get $\mathcal{X}\mathcal{Y} \in \mathcal{T}^{(2)}_{N_f+2G}$ (follows from Theorem 28 on page 23 of [36]). Therefore,

$$[\mathcal{X}\mathcal{Y}]_{\mathbf{g}} = 0 \qquad \forall \mathbf{g} \in \mathbb{Z}^2 \backslash \mathbb{Z}^2_{N_f+2G,\square}. \tag{4.17}$$

Now consider $[\mathcal{X}\mathcal{Y}]_{\mathbf{g}+N_f\mathbf{g}'}$ for $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$ and $\mathbf{0} \neq \mathbf{g}' \in \mathbb{Z}^2$. Since $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$, we have $|\mathbf{g}| \leq G$. And since $N_f = 4G+1$, it follows that $|(\mathbf{g}+\mathbf{g}'N_f)_i| > 3G+3$ for either $i = 1$ or $i = 2$. Therefore, $\mathbf{g}+\mathbf{g}'N_f \notin \mathbb{Z}^2_{N_f+2G,\square}$ and $[\mathcal{X}\mathcal{Y}]_{\mathbf{g}+\mathbf{g}'N_f} = 0$ by (4.17).

Therefore (4.16) implies that

$$[\mathrm{Q}_{N_f}(\mathcal{X}\mathcal{Y})]_{\mathbf{g}} = [\mathcal{X}\mathcal{Y}]_{\mathbf{g}+N_f\mathbf{0}} = [\mathcal{X}\mathcal{Y}]_{\mathbf{g}} \qquad \forall \mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$$

Part 3. This part follows directly from the definition of the algorithm and ideas discussed in Subsection 3.2.4, i.e. that a function in $\mathcal{T}^{(2)}_{N_f}$ can be represented as a matrix of nodal values or a matrix of Fourier coefficients and that the FFT and inverse FFT can be used to swap between these two representations. First, note that $\mathcal{Y}$ is represented in the matrix $\widehat{Y}$ with a matrix of Fourier coefficients before we pre-compute $Y \leftarrow \mathrm{ifft}(\widehat{Y})$ to represent $\mathcal{Y}$ with a matrix of nodal values.

Now consider what the algorithm does. Step 1 and 2 are equivalent to representing $\mathcal{X}$ with a matrix $\widehat{X}$ of Fourier coefficients. In Step 3, the representation of $\mathcal{X}$ is swapped to a matrix X of nodal values by computing the inverse FFT of $\widehat{X}$. In Step 4 we sample $\mathcal{X}\mathcal{Y}$ at nodal values and store the information in X. Sampling $\mathcal{X}\mathcal{Y}$ at these nodes corresponds to taking the Q_{N_f} projection of $\mathcal{X}\mathcal{Y}$. The matrix X is a representation of $\mathrm{Q}_{N_f}(\mathcal{X}\mathcal{Y})$ in terms of its nodal values. In Step 5 we swap the representation of

Memory Required for Implementation		
Part of Implementation	Amount	Type
Eigenvectors	$N_{EV} \times N$	double
Storage of A	$4 \times d \times N$	double
ARPACK	$4 \times N$ $2 \times N_{EV} \times N$	double double
PCG	$5 \times N$	double
Matrix-Vector Product	$2 \times d \times N$	complex double
Total	$(3N_{EV} + 8d + 9)N$	double

Table 4.1: Estimates for the memory required for the implementation of both the 1D and 2D Problems in terms of $N = \dim \mathrm{A}$, neglecting lower order terms. N_{EV} denotes the number of eigenpairs being sought.

$\mathrm{Q}_{N_f}(\mathcal{X}\mathcal{Y})$ to a matrix $\widehat{X}$ of Fourier coefficients by computing the FFT of X. In Step 6 we select the Fourier coefficients from X that correspond to $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$ This corresponds to taking $[\mathrm{Q}_{N_f}(\mathcal{X}\mathcal{Y})]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$. □

Now that we have an algorithm for computing matrix-vector products with A and we have specified that our implementation is using PCG and the IRA method, we present the total memory requirements of our implementation in Table 4.1. Note that we only worry about the leading order terms and we have ignored memory requirements that do not depend on $N = \dim \mathrm{A}$ and are generally small in comparison. Recall that $N = 2G + 1$ for the 1D problem and $N \leq 4G^2$ for the 2D problem.

Now we consider preconditioning A (where A is the matrix from (4.14)). The first preconditioner that we consider is the diagonal of A. Recall that $\mathrm{A} = \mathrm{D} - \mathrm{V}$ where D is a diagonal matrix and V is a dense matrix with entries

$$\mathrm{D}_{i(\mathbf{g}),i(\mathbf{g})} = |\boldsymbol{\xi} + 2\pi\mathbf{g}|^2 + \mathrm{K}$$
$$\mathrm{V}_{i(\mathbf{g}),i(\mathbf{g}')} = [\gamma]_{\mathbf{g}-\mathbf{g}'}$$

for $\mathbf{g}, \mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}}$. We define our preconditioner as

$$\mathrm{P} := \mathrm{diag}(\mathrm{A}) = \mathrm{D} - [\gamma]_0 \, \mathrm{I}$$

In practice we observe that using this preconditioner is optimal in the sense that PCG converges in $\mathcal{O}(1)$ iterations (independent of G). An informal explanation for this is that all of the contributions from the derivative components in the bilinear form of $a(\cdot, \cdot)$ are located in D and by preconditioning with the diagonal of A we negate their effect on the condition number of A.

We now prove two rigorous results about the condition number of $\mathrm{P}^{-1}\,\mathrm{A}$. First,

we prove a result for the 2D problem and then we prove a similar result for the 1D problem.

Theorem 4.21. *For any $C > 1$, if $\gamma \in PC'_p$ and*

$$\mathrm{K} \geq [\gamma]_0 + \tfrac{C+1}{C-1} 2^{11/4} F \sqrt{G} \qquad \textit{then} \qquad \kappa(\mathrm{P}^{-1}\,\mathrm{A}) \leq C$$

where F is a constant that depends on the discontinuities in $\gamma(\mathbf{x})$.

Note that we must choose $\mathrm{K} \to \infty$ as $G \to \infty$.

Proof. The proof of this result relies on Theorem 3.47 and Gershgorin's Circle Theorem which says: For any matrix T,

$$\sigma(\mathrm{T}) \subset \bigcup_{i=1}^{N} B(\mathrm{T}_{ii}, r_i)$$

where $B(\mathrm{T}_{ii}, r_i)$ is an open ball centred at T_{ii} with radius $r_i := \sum_{j \neq i}^{N} |\,\mathrm{T}_{ij}\,|$.

Our choice of P gives $(\mathrm{P}^{-1}\,\mathrm{A})_{i(\mathbf{g})i(\mathbf{g})} = 1$ for all $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$. We bound $r_{i(\mathbf{g})}$ in the following way. For $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$ we have

$$\begin{aligned}
r_{i(\mathbf{g})} &= \sum_{\substack{\mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}} \\ \mathbf{g}' \neq \mathbf{g}}} |(\mathrm{P}^{-1}\,\mathrm{A})_{i(\mathbf{g})i(\mathbf{g}')}| \leq \tfrac{1}{|\boldsymbol{\xi}+2\pi\mathbf{g}|^2+\mathrm{K}-[\gamma]_0} \sum_{\substack{\mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}} \\ \mathbf{g}' \neq \mathbf{g}}} |[\gamma]_{\mathbf{g}-\mathbf{g}'}| \\
&\leq \tfrac{1}{\mathrm{K}-[\gamma]_0} \sum_{\substack{\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}} \\ \mathbf{g} \neq 0}} |[\gamma]_{\mathbf{g}}| \qquad \leq \tfrac{1}{\mathrm{K}-[\gamma]_0} \sum_{\substack{|g_1|+|g_2| \leq 2\sqrt{2}G \\ \mathbf{g} \neq 0}} |[\gamma]_{\mathbf{g}}| \\
&= \tfrac{1}{\mathrm{K}-[\gamma]_0} \sum_{n=1}^{\lfloor 2\sqrt{2}G \rfloor} \sum_{|g_1|+|g_2|=n} |[\gamma]_{\mathbf{g}}| \\
&\leq \tfrac{1}{\mathrm{K}-[\gamma]_0} \sum_{n=1}^{\lfloor 2\sqrt{2}G \rfloor} \left(\sum_{|g_1|+|g_2|=n} 1 \right)^{\frac{1}{2}} \left(\sum_{|g_1|+|g_2|=n} |[\gamma]_{\mathbf{g}}|^2 \right)^{\frac{1}{2}} \quad \text{by Cauchy-Schwarz} \\
&= \tfrac{1}{\mathrm{K}-[\gamma]_0} \sum_{n=1}^{\lfloor 2\sqrt{2}G \rfloor} (4n)^{1/2}\, C_n \qquad \text{where } C_n := \left(\textstyle\sum_{|g_1|+|g_2|=n} |[\gamma]_n|^2 \right)^{1/2} \\
&\leq \tfrac{2F}{\mathrm{K}-[\gamma]_0} \sum_{n=1}^{\lfloor 2\sqrt{2}G \rfloor} n^{-1/2} \qquad \text{since } C_n \leq F n^{-1} \text{ by Theorem 3.47} \\
&\leq \tfrac{2F}{\mathrm{K}-[\gamma]_0} \left(1 + \int_1^{2\sqrt{2}G} x^{-1/2} dx \right) \qquad \text{by Lemma 3.9} \\
&\leq \tfrac{2^{11/4} F \sqrt{G}}{\mathrm{K}-[\gamma]_0} \leq \tfrac{C-1}{C+1} \qquad \text{if } \mathrm{K} \geq [\gamma]_0 + \tfrac{C+1}{C-1} 2^{11/4} F \sqrt{G}
\end{aligned}$$

Note that F depends on the number and height of the discontinuities in $\gamma(\mathbf{x})$.

Applying Gershgorin's Circle Theorem we get

$$\sigma(\mathrm{P}^{-1}\,\mathrm{A}) \subset \left[1 - \tfrac{C-1}{C+1}, 1 + \tfrac{C-1}{C+1}\right].$$

Therefore $\kappa(\mathrm{P}^{-1}\,\mathrm{A}) = \frac{\lambda_{max}}{\lambda_{min}} \leq C$. □

Now we present the corresponding 1D result for diagonal preconditioning.

Theorem 4.22. *Let* A *be the matrix from* (4.14) *corresponding to the 1D problem. That is,*

$$\mathrm{A} = \mathrm{D} - \mathrm{V}$$

where D *is a diagonal matrix and* V *is a Toeplitz matrix with entries given by*

$$\mathrm{D}_{ii} = (\xi^2 + 2\pi(i - G - 1))^2 + \mathrm{K}$$
$$\mathrm{V}_{ij} = [\gamma]_{i-j}$$

for $i, j = 1, ..., N = 2G + 1$. *Define a preconditioner*

$$\mathrm{P} := \mathrm{diag}(\mathrm{A}) = \mathrm{D} - [\gamma]_0\,\mathrm{I}$$

Then for any $C > 1$, *if*

$$\mathrm{K} \geq [\gamma]_0 + \tfrac{C+1}{C-1} 2F(1 + \log G) \qquad \textit{then} \qquad \kappa(\mathrm{P}^{-1}\,\mathrm{A}) \leq C.$$

F *is a constant that depends on* γ.

Proof. This proof is similar to the proof of Theorem 4.21 and we again use Gershgorin's Circle Theorem. With our definition of P we get $(\mathrm{P}^{-1}\,\mathrm{A})_{ii} = 1$ for all $i = 1, \ldots, N$. We then bound r_i in the following way

$$\begin{aligned}
r_i &= \sum_{i \neq j \in \mathbb{Z}^1_{G,\mathrm{o}}} |(\mathrm{P}^{-1}\,\mathrm{A})_{ij}| \leq \tfrac{1}{(\xi + 2\pi(i-G-1))^2 + \mathrm{K} - [\gamma]_0} \sum_{i \neq j \in \mathbb{Z}^1_{G,\mathrm{o}}} |[\gamma]_{i-j}| \\
&\leq \tfrac{1}{\mathrm{K} - [\gamma]_0} \sum_{0 \neq |j| \leq G} |[\gamma]_j| \\
&\leq \tfrac{2F}{\mathrm{K} - [\gamma]_0} \sum_{n=1}^{G} n^{-1} \qquad \text{since } |[\gamma]_n| \leq F|n|^{-1} \text{ by Lemma 3.41} \\
&\leq \tfrac{2F}{\mathrm{K} - [\gamma]_0} \left(1 + \int_1^G x^{-1} dx\right) \qquad \text{by Lemma 3.9} \\
&= \tfrac{2F(1+\log G)}{\mathrm{K} - [\gamma]_0} \\
&\leq \tfrac{C-1}{C+1} \qquad \text{if } \mathrm{K} \geq [\gamma]_0 + \tfrac{C+1}{C-1} 2F(1 + \log G)
\end{aligned}$$

Applying Gershgorin's Circle Theorem we get

$$\sigma(\mathrm{P}^{-1}\,\mathrm{A}) \subset \left[1 - \tfrac{C-1}{C+1}, 1 + \tfrac{C-1}{C+1}\right].$$

Therefore $\kappa(\mathrm{P}^{-1}\,\mathrm{A}) = \frac{\lambda_{max}}{\lambda_{min}} \leq C$. □

Theorems 4.21 and 4.22 imply that we should choose a sufficiently large shift K that depends on G and precondition with the diagonal of A. However, practice tells us that choosing a large K results in more iterations for the IRA method to converge. An explanation for this follows from the fact that as we increase K the relative distance between the eigenvalues of A (and A^{-1}) decreases and this has a negative effect on the performance of our our eigensolver, see Theorem 3.82. Also, if K is very large then we might experience round-off errors when shifting back and calculating $\beta^2 = -(\lambda - \mathrm{K})$.

Instead of preconditioning with the diagonal of A with K large, we choose K just large enough to satisfy Lemma 4.7 and precondition with the following block matrix (in the 2D case)

$$\mathrm{P} = \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix} \tag{4.18}$$

where B_1 is a $N_b \times N_b$ dense matrix with entries that are the same as the entries in A, and B_2 is a $(N - N_b) \times (N - N_b)$ diagonal matrix that has diagonal entries that correspond to the diagonal of A, i.e.

$$\begin{aligned} (B_1)_{ij} &= \mathrm{A}_{ij} && \text{for } i,j = 1,\ldots,N_b \\ (B_2)_{ii} &= \mathrm{A}_{(i+N_b,i+N_b)} && \text{for } i = 1,\ldots,(N-N_b). \end{aligned}$$

This choice of preconditioner keeps the advantages of preconditioning with the diagonal of A as well as picking the parts of A that correspond to the low frequency plane wave terms. This is because the block B_1 corresponds to the entries of A that are generated from the N_b basis functions with smallest frequency, i.e. the $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$ with smallest $|\mathbf{g}|$.

An important property for a preconditioner is that we can compute the action of P^{-1} easily. In this case if we can compute the action of B_1^{-1} and B_2^{-1} then we can compute the action of P^{-1}. B_2^{-1} is trivial since B_2 is a diagonal matrix. To compute the action of B_1^{-1} we solve a linear system using Cholesky factorization and back substitution at a cost of $\mathcal{O}(N_b^3)$ operations for the Cholesky factorization and $\mathcal{O}(N_b^2)$ operations for the back substitution. In practice, we compute the Cholesky factorization only once and store the factors.

Other than choosing $N_b \leq N$, we are free to tune our preconditioner by choosing N_b to give us the best results. The larger we choose N_b the more information from A is represented in P. Therefore, we expect $\mathrm{P}^{-1}\,\mathrm{A}$ to more closely approximate the

identity matrix and have a small condition number. However, the cost of computing P^{-1} increases with large N_b. In practice, we can choose N_b up to 1000.

In the 1D case, the structure of A is slightly different because the ordering of the entries is different. The entries of A that correspond to the low frequency basis functions are located in the middle of the matrix, and not the top left corner. Therefore, in the 1D case we choose our preconditioner to be

$$\mathrm{P} = \begin{bmatrix} B_2 & 0 & 0 \\ 0 & B_1 & 0 \\ 0 & 0 & B_3 \end{bmatrix} \tag{4.19}$$

where B_1 is a $(2N_b + 1) \times (2N_b + 1)$ dense matrix with entries that correspond to the same entries in A, and B_2 and B_3 are $(N - N_b) \times (N - N_b)$ diagonal matrices with entries on the diagonal that correspond to the diagonal of A, i.e.

$$\begin{aligned} (B_1)_{ij} &= \mathrm{A}_{(i+(G-N_b),j+(G-N_b))} && \text{for } i,j = 1,\ldots,2N_b+1 \\ (B_2)_{ii} &= \mathrm{A}_{ii} && \text{for } i = 1,\ldots,G-N_b \\ (B_3)_{ii} &= \mathrm{A}_{(i+(G+1+N_b),i+(G+1+N_b))} && \text{for } i = 1,\ldots,G-N_b. \end{aligned}$$

Now we must choose N_b so that $1 \leq N_b \leq G$. In practice, we choose N_b up to 500 for the 1D case.

For both the 1D and 2D problems we observe that this new preconditioner is optimal in the sense that we get convergence in $\mathcal{O}(1)$ iterations in the PCG algorithm.

Now we will consider the computing requirements of our implementation for Model Problems 1 - 4 that we defined in Section 4.1.7. As we will see in Subsection 4.2.4, the computing requirements are the most extreme when we compute reference solutions and we give a summary of the parameters, memory and CPU time requirements for these problems in Table 4.2. All of the computations in this thesis were carried out on a Dual Core AMD Opteron Processor 285 with speed 2600 MHz and 1024 Kb cache, and 8 Gb of memory. All of the programs were written in Fortran 95 and compiled with GNU Fortran 4.2.0. Other libraries that were used include: LAPACK 3.1.1-4, BLAS 3.1.1-4, ARPACK 2.1-7 and FFTW 4.2-3.1.2-1.

Finally, in Tables 4.3 and 4.4 and Figure 4-5 we present data that confirm the claims that we have made throughout this subsection.

In Table 4.3 we have solved Model Problem 2 (from Section 4.1.7) using different preconditioners and varying G (and a shift $\mathrm{K} = \gamma_g + \pi^2 + \frac{1}{2}$ unless otherwise stated). The different precondtioners are defined as $\mathrm{P}_1 = \mathrm{I}$, $\mathrm{P}_2 = \mathrm{diag(A)}$, $\mathrm{P}_3 = \mathrm{diag(A)}$ (with large shift $\mathrm{K} = 5000$) and $\mathrm{P}_4 = \mathrm{P}$ from (4.19) (where $N_b = 2^{k-1}$ for $k \leq 9$ and $N_b = 2^9$ for $k \geq 10$). We have recorded the number of iterations that PCG requires per IRA iteration as well as the number of restarts that IRA needs. The total number of calls

Computing Reference Solutions to Model Problems 1-4				
Model Problem	1	2	3	4
N_{EV} (# of eigenpairs)	5	60	5	30
G	$2^{18}-1$	$2^{18}-1$	$2^{10}-1$	$2^{10}-1$
$N = \dim \mathrm{A}$	$\approx 5 \times 10^5$	$\approx 5 \times 10^5$	$\approx 3 \times 10^6$	$\approx 3 \times 10^6$
$(N_f)^d$ (FFT size)	2^{20}	2^{20}	2^{24}	2^{24}
Total Memory (Mb)	≈ 130	≈ 750	≈ 1000	≈ 2500
CPU time (seconds)	$\mathcal{O}(10^2)$	$\mathcal{O}(10^3)$	$\mathcal{O}(10^3)$	$\mathcal{O}(10^4)$

Table 4.2: The details of the largest problems that we solve when we compute the reference solutions for Model Problems 1-4 in Subsection 4.2.4.

to PCG required by IRA is approximately (number of restarts)$\times N_{EV}$ since we have set IRA to restart after N_{EV} iterations if it has not already converged (recall N_{EV} denotes the number of eigenpairs being sought).

Table 4.4 is similar to Table 4.3 except it is for solving Model Problem 4 instead of Model Problem 2. For this table, $\mathrm{P}_4 = \mathrm{P}$ from (4.18) (with $N_b = 2^k$ for $k \leq 5$ and $N_b = 2^9$ for $k \geq 6$).

In these two tables we see that the number of iterations required by PCG is $\mathcal{O}(1)$ when we use the diagonal of A as a preconditioner and that even fewer iterations are needed by PCG when K is large. However, choosing K large has an adverse effect on the number of iterations required by our eigensolver. We see that it is possible to get the best of both worlds using the preconditioner that we defined in (4.18) and (4.19). Note that the results for Model Problems 2 and 4 are also representative of the results for Model Problems 1 and 3.

In Figure 4-5 we have plotted the CPU time required to solve Model Problems 1-4 for varying $N = \dim \mathrm{A}$ using the preconditioner P_4. The plots confirms the overarching claim that the total implementation only requires $\mathcal{O}(N \log N)$ operations. Note that the kinks in the Model Problem 1 and 2 lines are due to how we choose N_b in the preconditioner.

In conclusion we have a very efficient algorithm for computing matrix-vector products for both the 1D and 2D problems using FFT, we observe that we have an optimal preconditioner that allows us to solve linear systems in a fixed number of iterations independent of the size of the system, and we have an iterative Krylov subspace eigensolver that also converges in a fixed number of iterations independent of the system size. Therefore, we have an implementation that solves (4.14) in $\mathcal{O}(N \log N)$ operations. This is in contrast to a direct method that would require $\mathcal{O}(N^3)$ iterations. (Recall that in 2D $N = \mathcal{O}(G^2)$ and in 1D $N = 2G + 1$).

Model Problem 2 with different preconditioners								
$G = 2^k - 1$	PCG iterations				IRA restarts			
k	P_1	P_2	P_3	P_4	P_1	P_2	P_3	P_4
6	16	26	8	17	2	2	2	2
7	45	25	8	12	2	2	5	2
8	98	25	8	9	2	2	8	2
9	X	25	8	7	X	2	10	2
10	X	25	8	6	X	2	10	2
11	X	25	8	6	X	2	10	2
12	X	25	8	6	X	2	10	2

Table 4.3: Solving Model Problem 2 with different preconditioners and varying G (with shift $\mathrm{K} = \gamma_g + \pi^2 + \frac{1}{2}$ unless otherwise stated).

Model Problem 4 with different preconditioners								
$G = 2^k - 1$	PCG iterations				IRA restarts			
k	P_1	P_2	P_3	P_4	P_1	P_2	P_3	P_4
3	28	36	8	36	6	6	11	6
4	50	38	8	39	7	7	22	11
5	99	38	8	39	7	7	41	7
6	204	39	8	18	7	7	65	7
7	410	39	8	18	7	7	96	7

Table 4.4: Solving Model Problem 4 with different preconditioners and varying G (with shift $\mathrm{K} = \gamma_g + \pi^2 + \frac{1}{2}$ unless otherwise stated).

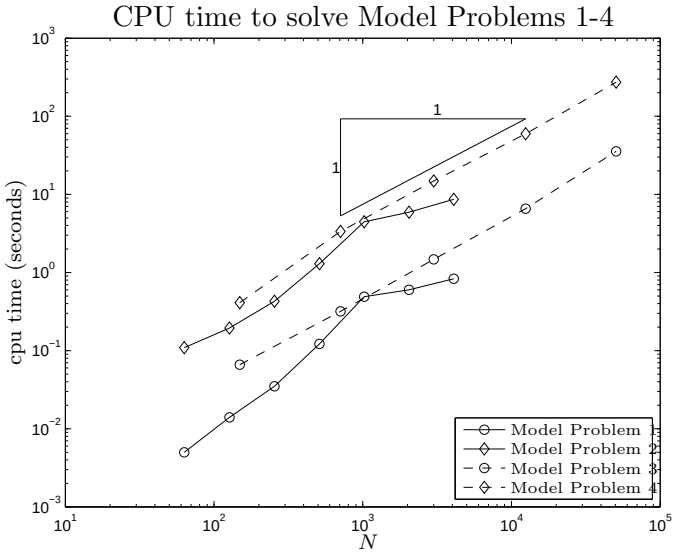

Figure 4-5: Plot of CPU time vs. $N = \dim \mathrm{A}$ required to solve Model Problems 1-4 using the preconditioner P_4.

4.2.3 Error Analysis

In this subsection we derive error bounds for the eigenvalue and eigenfunction errors for the approximate solution to Problem 4.6 that we obtain by solving Problem 4.17, i.e. by applying the spectral Galerkin method to Problem 4.6. The error bounds are derived so that we can see the rate at which the errors decrease as we increase G. That is, as we include more basis functions in our finite dimensional space $\mathcal{S}_G$ (see (4.8)), what reduction in the errors should we expect to see in our numerical computations?

These results are based on results in Section 3.5 and are an application of [6]. The main analytical tool that we use is the solution operator for Problem 4.6, T, which was defined in Subsection 4.1.4. Problem 4.17 also has a solution operator, T_G (defined in a similar way to T_n in (3.43)).

We will predominantly focus on the 2D problem in this subsection, however, all of the results also apply to the 1D problem with very similar proofs. At the end of this subsection we present an additional result that only applies to the 1D problem.

We begin by examining the properties of T_G. The following lemma proves that T_G has similar properties to those of T (see Lemma 4.9) as well as proving that $\mathrm{T}_G \to \mathrm{T}$ in norm as $G \to \infty$. We also prove an approximation error bound in the subspace $\mathcal{S}_G$ for approximating eigenfunctions of Problem 4.6. The results in the following lemma are all needed for the main theorem of this section.

Lemma 4.23. *Let $\gamma \in PC_p$. Then the following properties hold for* T, T_G *and* $\mathcal{S}_G$.

1. $\mathrm{T}_G = \mathrm{P}_G\mathrm{T}$ *where* P_G *is the projection from* H^1_p *onto* $\mathcal{S}_G$ *defined by*

$$a(\mathrm{P}_G u - u, v) = 0 \qquad \forall u \in H^1_p, \forall v \in \mathcal{S}_G.$$

2. $\mathrm{T}_G : H^1_p \to H^1_p$ *is a bounded, compact, self-adjoint operator with respect to* $a(\cdot,\cdot)$.

3. *For* $u \in H^1_p$ *and* $\epsilon > 0$,

$$\inf_{\chi \in \mathcal{S}_G} \| \mathrm{T}\, u - \chi \|_{H^1_p} \lesssim G^{-3/2+\epsilon} \|u\|_{H^1_p}.$$

4. *For* $\epsilon > 0$,

$$\| \mathrm{T} - \mathrm{T}_G \|_{H^1_p} \lesssim G^{-3/2+\epsilon}.$$

5. *If* u *is an eigenfunction of Problem 4.6 then, for* $\epsilon > 0$,

$$\inf_{\chi \in \mathcal{S}_G} \|u - \chi\|_{H^1_p} \lesssim G^{-3/2+\epsilon} \|u\|_{H^1_p}.$$

Proof. Part 1 is Part 1 of Lemma 3.74 with $\mathcal{S}_n = \mathcal{S}_G$.
Part 2. T_G is bounded since $\mathrm{T}_G = \mathrm{P}_G\,\mathrm{T}$ from Part 1 and P_G and T are both bounded.

T_G compact follows from Part 1 since P_G is bounded and linear and T is compact (Lemma 4.9 and the fact that the composition of a compact operator with a linear bounded operator is compact). T_G is self-adjoint by the same argument as for T self-adjoint (see Lemma 4.9).
Part 3. With $\mathrm{P}_G^{(\mathcal{S})}$ defined in Subsection 3.2.5,

$$\begin{aligned}
\inf_{\chi\in\mathcal{S}_G} \|\,\mathrm{T}\,u-\chi\|_{H_p^1} &\le \|\,\mathrm{T}\,u-\mathrm{P}_G^{(\mathcal{S})}\,\mathrm{T}\,u\|_{H_p^1} && \text{choosing } \chi=\mathrm{P}_G^{(\mathcal{S})}\,\mathrm{T}u\\
&\le G^{-3/2+\epsilon}\|\,\mathrm{T}\,u\|_{H_p^{5/2-\epsilon}} && \text{by Lemma 3.30}\\
&\lesssim G^{-3/2+\epsilon}\|u\|_{H_p^1} && \text{by Theorem 4.11.}
\end{aligned}$$

Part 4 follows from Part 3 using Part 2 of Lemma 3.74,

$$\begin{aligned}
\|\,\mathrm{T}_G-\mathrm{T}\,\|_{H_p^1} &= \sup_{u\in H_p^1}\frac{\|\,\mathrm{T}_G u-\mathrm{T}u\|_{H_p^1}}{\|u\|_{H_p^1}}\\
&\lesssim \sup_{u\in H_p^1}\inf_{\chi\in\mathcal{S}_G}\frac{\|\,\mathrm{T}u-\chi\|_{H_p^1}}{\|u\|_{H_p^1}} && \text{by Part 2 of Lemma 3.74}\\
&\lesssim G^{-3/2+\epsilon} && \text{by Part 3.}
\end{aligned}$$

Part 5 uses the same argument as Part 3.

$$\begin{aligned}
\inf_{\chi\in\mathcal{S}_G}\|u-\chi\|_{H_p^1} &\le \|u-\mathrm{P}_G^{(\mathcal{S})}\,u\|_{H_p^1}\\
&\le G^{-3/2+\epsilon}\|u\|_{H_p^{5/2-\epsilon}} && \text{by Lemma 3.30}\\
&\lesssim G^{-3/2+\epsilon}\|u\|_{H_p^1} && \text{by Theorem 4.11.}
\end{aligned}$$

□

We can now apply the theory in [6] by using Theorem 3.68 to obtain our main theorem for this section.

Theorem 4.24. *Let $\gamma\in PC_p$ and let λ be an eigenvalue of Problem 4.6 with multiplicity m and corresponding eigenspace M. Then for sufficiently large G and arbitrarily small $\epsilon>0$, there exist m eigenvalues $\lambda_1(G),\dots,\lambda_m(G)$ of Problem 4.17 (counted according to their multiplicty) with corresponding eigenspaces $M_1(\lambda_1),\dots,M_m(\lambda_m)$ and*

$$\mathcal{M}_G := \bigoplus_{j=1}^{m} M_j(\lambda_j)$$

such that

$$\delta(M,\mathcal{M}_G)\lesssim G^{-3/2+\epsilon}$$

and

$$|\lambda - \lambda_j| \lesssim G^{-3+2\epsilon} \qquad \textit{for } j = 1, \ldots, m.$$

Here, $\delta(\cdot,\cdot)$ is defined as in Definition 3.64 but with $\mathcal{H} = H_p^1$ since all of the eigenspaces are subspaces of H_p^1.

Proof. The proof of this result is a direct application Theorem 3.68 and Lemma 3.71. We first check that the assumptions of Theorem 3.68 are satisfied.

1. Our Hilbert space is $H_p^1(\Omega)$ and $a(\cdot,\cdot)$ is an inner product for this Hilbert space by Corollary 4.8.

2. T is bounded, compact and self-adjoint on this Hilbert space by Lemma 4.9.

3. T_G (for $G \in \mathbb{N}$) are a family of bounded, compact operators such that $\mathrm{T}_G \to \mathrm{T}$ in norm as $G \to \infty$ by Lemma 4.23.

4. $\frac{1}{\lambda}$ is an eigenvalue of T with eigenspace M by Lemma 3.71.

This completes checking the assumptions of Theorem 3.68. Applying Theorem 3.68 we get

$$\delta(M, \mathcal{M}_G) \lesssim \|(\mathrm{T} - \mathrm{T}_G)|_M\|_{H_p^1}$$

and

$$|\lambda - \lambda_j| \lesssim \sum_{i,k=1}^{m} |a((\mathrm{T} - \mathrm{T}_G)\phi_i, \phi_k)| + \|(\mathrm{T} - \mathrm{T}_G)|_M\|_{H_p^1}^2 \qquad j = 1, \ldots, m$$

where $\phi_1, \ldots, \phi_m$ is a basis for M.

The result follows using Lemma 3.74 and Parts 3-5 of Lemma 4.23. □

In the special case of the 1D problem we can improve these bounds so that we may choose $\epsilon = 0$. This is based on being able to derive an improved approximation error result and we present this now.

Lemma 4.25. *In 1D, let $u \in H_p^1$. Then*

$$\inf_{\chi \in S_G} \|\mathrm{T}\, u - \chi\|_{H_p^1} \lesssim G^{-3/2}$$

Proof. Since $u \in H_p^1$, by Theorem 4.16 we know that $\mathrm{T}\, u$ and $(\mathrm{T}\, u)'$ are absolutely continuous and $(\mathrm{T}\, u)''$ is continuous except where $\gamma(x)$ is discontinuous and is absolutely continuous on the intervals of continuity. Theorem 39 on page 26 of [36] then implies that $[(\mathrm{T}\, u)'']_g = \mathcal{O}(g^{-1})$. Since $[(\mathrm{T}\, u)'']_g = (i2\pi g)^2 [\mathrm{T}\, u]_g$ for all $g \in \mathbb{Z}$ we then get

$[\mathrm{T}\,u]_g = \mathcal{O}(g^{-3})$ and

$$\begin{aligned}
\inf_{\chi \in S_G} \|\,\mathrm{T}\,u - \chi\|_{H_p^1}^2 &\leq \|\,\mathrm{T}\,u - \mathrm{P}_G^{(\mathcal{S})}\,\mathrm{T}\,u\|_{H_p^1}^2 \\
&= \sum_{|g|>G} |g|^2 |[\mathrm{T}\,u]_g|^2 \\
&\lesssim \sum_{g=G+1}^{\infty} g^{-4} && \text{since } [\mathrm{T}\,u]_g = \mathcal{O}(g^{-3}) \\
&\leq \int_G^{\infty} x^{-4} dx && \text{by Lemma 3.9} \\
&= \tfrac{1}{3} G^{-3}
\end{aligned}$$

The result follows by taking the square root of both sides. □

The approximation error of an eigenfunction of the 1D problem can also be bounded using the same technique. We can then obtain the results from Theorem 4.24 with $\epsilon = 0$ by the same proof, using Lemma 4.25 instead of Parts 3-5 of Lemma 4.23.

To recap, we have proven that Problem 4.17 approximates Problem 4.6 in the sense that given an eigenpair of Problem 4.6 and sufficiently large G, then there is an eigenpair of Problem 4.17 that approximates the eigenpair of Problem 4.6. We have proven error bounds for the eigenvalue and eigenfunction error in terms of G. We can now say that as G gets bigger we know that the eigenvalue and eigenfunction errors will decrease at specific rates. Moreover, the results for the H_p^1 error of the eigenfunctions decreases at an optimal rate with respect to G since our eigenfunction error results are in terms of the approximation error for $\mathcal{S}_G$ in H_p^1. This means that the eigenfunction error is equivalent to the error between the exact eigenfunction and the best possible approximation of that eigenfunction from $\mathcal{S}_G$.

Interestingly, our theory implies that the convergence of the eigenvalues is twice as fast as the convergence of the eigenfunctions. This result is analogous to the convergence of numerical linear algebra techniques for solving symmetric matrix eigenproblems where the convergence of eigenvalues is twice as fast as the convergence of eigenvectors.

We must also point out that the convergence of this method is *not* superalgebraic. We can not expect superalgebraic convergence (despite having global basis functions) because the eigenfunctions of Problem 4.6 are not in C_p^{∞}.

The next subsection will verify the results of this subsection with some numerical experiments.

4.2.4 Examples

In this section we solve (4.14) for Model Problems 1-4 (see Section 4.1.7) for increasing values of G to see how the eigenvalues and eigenfunctions of these problems converge. In particular, we would like to verify our error estimates from Theorem 4.24.

We compare the eigenvalues and eigenfunctions of (4.14) with a reference solution that has been computed with an especially large value of G (Model Problems 1 and 2: $G = 2^{18} - 1$ which corresponds to $N_f = 2^{20}$; Model Problems 3 and 4: $G = 2^{10} - 1$ corresponding to $N_f = 2^{12}$). We calculate the relative error of eigenvalues and the H_p^1 norm of eigenfunction errors. All of the plots will have logarithmically scaled axes so that a function $y = Cx^r$ with constants C and r will be represented as a straight line of slope r on a plot with horizontal axis x and vertical axis y. Our analysis has focused on obtaining the correct rate of convergence and so we are interested in the slope of the lines we plot.

We see that in Figures 4-6 to 4-9 the eigenfunction errors decay with $\mathcal{O}(G^{-3/2})$ while the eigenvalue errors decay with $\mathcal{O}(-3)$. Both of these rates agree with the error bounds that we proved in Theorem 4.24 for both the 1D and 2D problems. Moreover, it appears that the hidden constant in the error bounds of Theorem 4.24 does not depend on ϵ.

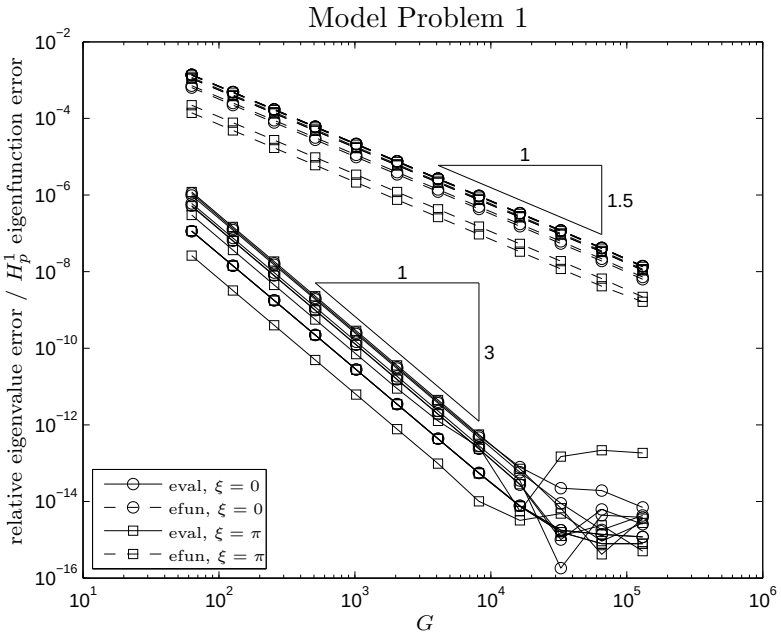

Figure 4-6: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. G for the first 5 eigenpairs of Model Problem 1 (solved for both $\xi = 0$ and $\xi = \pi$).

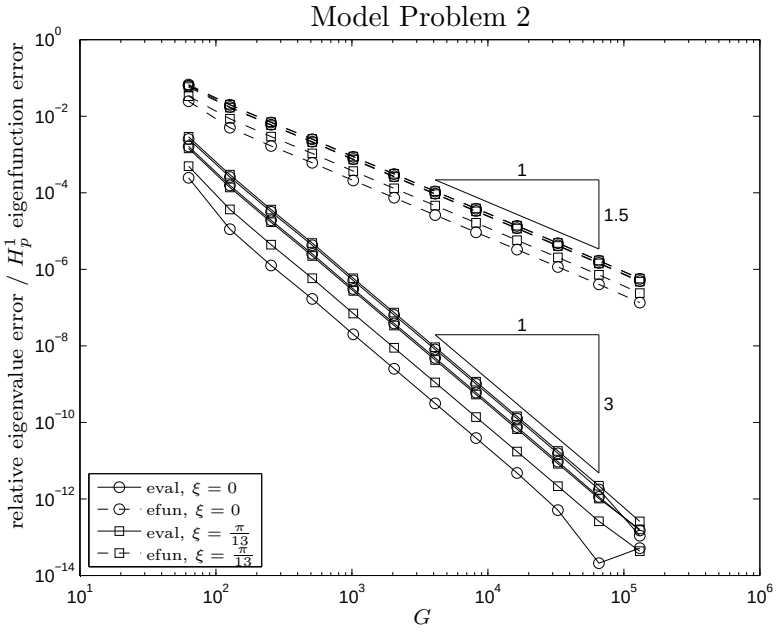

Figure 4-7: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. G for the 37-39th eigenpairs of Model Problem 2 (solved for both $\xi = 0$ and $\xi = \frac{\pi}{13}$).

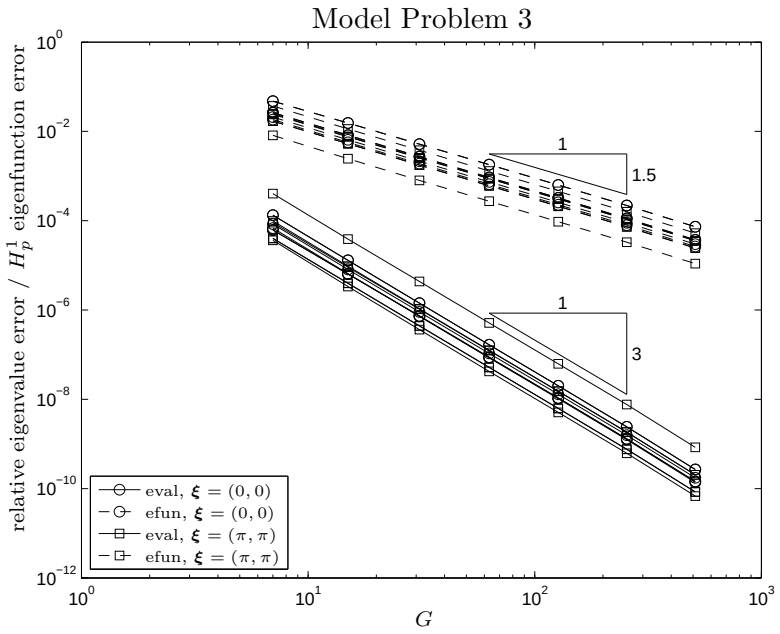

Figure 4-8: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. G for the first 5 eigenpairs of Model Problem 3 (solved for both $\boldsymbol{\xi} = (0,0)$ and $\boldsymbol{\xi} = (\pi,\pi)$).

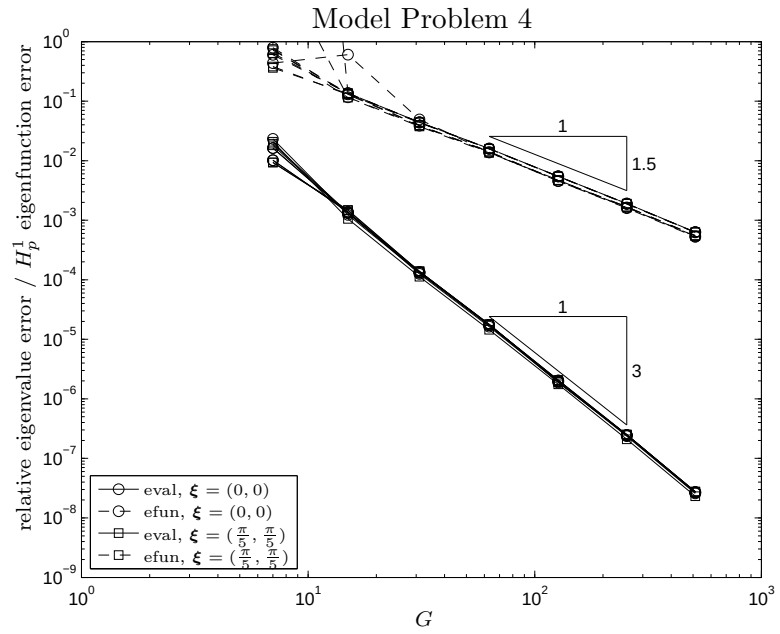

Figure 4-9: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. G for the 23-27th eigenpairs of Model Problem 4 (solved for both $\boldsymbol{\xi} = (0,0)$ and $\boldsymbol{\xi} = (\frac{\pi}{5},\frac{\pi}{5})$).

4.3 Smoothing

In the previous section we applied a standard spectral Galerkin method to Problem 4.6. If we ignored the fact that $\gamma(\mathbf{x})$ is discontinuous, then we might have expected superalgebraic convergence since the method has global basis functions. However, we saw that the eigenfunctions of Problem 4.6 are not C^∞ and therefore, we could only obtain algebraic convergence of limited order. Methods that attempt to recover faster (possibly superalgebraic) convergence have been suggested in [40], [53], [62], [63], [64] and [66]. All of the methods require that an *effective* n^2 that is smooth is used instead of a discontinuous n^2. In this thesis we focus on the method used in [62], [63], [64] and [66]. The method first modifies the operator (4.2) so that $\gamma(\mathbf{x})$ is a smooth function and then the same spectral Galerkin method is applied. In this section we examine the convergence properties of this method.

This section is divided into the three subsections. In the first subsection we define the new method. This is done by first defining the infinite dimensional *smooth* problem and then approximating the solution to this smooth problem via the spectral Galerkin method. In the second subsection we derive error bounds for the errors of this new method. The error is split into the error between the original problem and the smooth problem and the error from applying the spectral Galerkin method to the smooth problem. To obtain bounds for these errors it will be necessary to prove some properties of the smooth problem and this is included in the second subsection. Finally, in the third subsection we present some examples that verify our theoretical results.

In this section we assume that $\gamma \in PC'_p$ (see Definition 3.37). We make this assumption so that we can apply Theorem 3.47.

4.3.1 The method

In this subsection we define the new method as well as some properties that will be useful in the rest of this section. Let $\mathcal{G}(\mathbf{x})$ be a normalized Gaussian function defined by

$$\mathcal{G}(\mathbf{x}) = C_{\mathcal{G}} \exp\left(-\frac{|\mathbf{x}|^2}{2\Delta^2}\right) \tag{4.20}$$

for small $\Delta > 0$. In the 2D problem the normalization constant is $C_{\mathcal{G}} = \frac{1}{2\pi\Delta^2}$ and in the 1D problem the normalization constant is $C_{\mathcal{G}} = \frac{1}{\sqrt{2\pi}\Delta}$. The parameter Δ determines the "effective" width of the Gaussian function, and as $\Delta \to 0$, $\mathcal{G}$ approaches the Dirac delta function. In the papers where this method is used Δ is referred to as FWHM (Full-Width-Half-Maximum). Using this Gaussian function we smooth the piecewise constant coefficient function $\gamma(\mathbf{x})$ and define $\widetilde{\gamma}(\mathbf{x})$ as

$$\widetilde{\gamma}(\mathbf{x}) := (\mathcal{G} * \gamma)(\mathbf{x}) = \int_{\mathbb{R}^d} \mathcal{G}(\mathbf{x}-\mathbf{y})\gamma(\mathbf{y})d\mathbf{x}.$$

Now Δ determines the amount of smoothing. Large Δ corresponds to a lot of smoothing while $\Delta = 0$ corresponds to no smoothing provided we consider $\mathcal{G}$ in the distributional sense. See Figure 4-10 for an example of $\widetilde{\gamma}(x)$ for Model Problem 1 (see Section 4.1.7). Before we define the smooth problem let us state a result about $\widetilde{\gamma}(\mathbf{x})$ and its relationship to $\gamma(\mathbf{x})$.

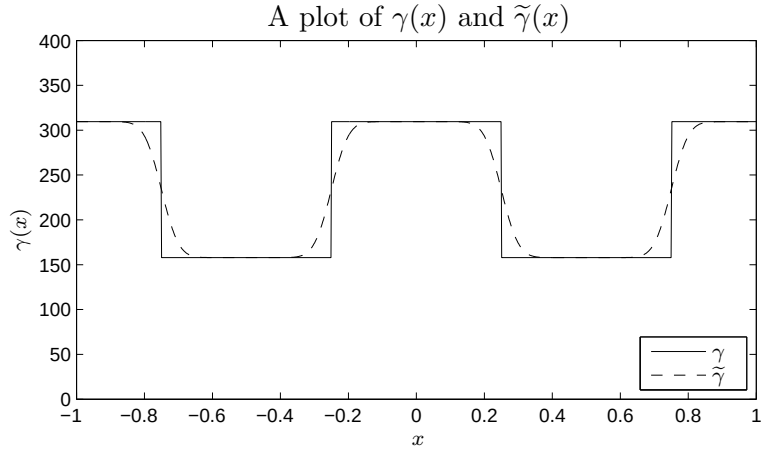

Figure 4-10: Plot of $\widetilde{\gamma}(x)$ in 1D.

Lemma 4.26. *With $\gamma \in PC'_p$ and $\widetilde{\gamma}(\mathbf{x})$ defined above, $s \in \mathbb{R}$ and $\Delta > 0$ the following three properties hold*

1. *The Fourier coefficients of $\widetilde{\gamma}(\mathbf{x})$ are related to the Fourier coefficients of $\gamma(\mathbf{x})$ by*

$$[\widetilde{\gamma}]_{\mathbf{g}} = \mathrm{e}^{-2\pi^2|\mathbf{g}|^2\Delta^2}[\gamma]_{\mathbf{g}} \qquad \forall \mathbf{g} \in \mathbb{Z}^2$$

2. $$\|\gamma - \widetilde{\gamma}\|_{H^s_p} \lesssim \Delta^{-s+1/2} \qquad -\tfrac{3}{2} < s < \tfrac{1}{2}$$

3. $$\|\widetilde{\gamma}\|_{H^s_p} \lesssim \begin{cases} \Delta^{-s+1/2} & s > \frac{1}{2} \\ \sqrt{\log(\Delta^{-1})} & s = \frac{1}{2} \\ 1 & s < \frac{1}{2} \end{cases}$$

Proof. Part 1. In this proof we will need the following.

$$\begin{aligned} \int_{\mathbb{R}} \exp\left(-\tfrac{y^2}{2\Delta^2} - i2\pi n y\right) dy &= \int_{\mathbb{R}} \exp\left(-\tfrac{(y+i2\pi n\Delta^2)^2}{2\Delta^2} - 2\pi^2 n^2 \Delta^2\right) dy \\ &= \mathrm{e}^{-2\pi^2 n^2\Delta^2} \int_{\mathbb{R}} \exp\left(-\tfrac{\eta^2}{2\Delta^2}\right) d\eta \\ &= \sqrt{2}\Delta\, \mathrm{e}^{-2\pi^2 n^2\Delta^2} \int_{\mathbb{R}} \mathrm{e}^{-\tau^2}\, d\tau = \sqrt{2\pi}\Delta\, \mathrm{e}^{-2\pi^2 n^2\Delta^2}. \end{aligned} \tag{4.21}$$

Using (4.21), for $\mathbf{g} \in \mathbb{Z}^2$, we get

$$\begin{aligned}
[\widetilde{\gamma}]_{\mathbf{g}} &= \int_\Omega \widetilde{\gamma}(\mathbf{x})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \\
&= \int_\Omega \left(\int_{\mathbb{R}^2} \mathcal{G}(\mathbf{y})\gamma(\mathbf{x}-\mathbf{y}) d\mathbf{y} \right) \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \\
&= \int_\Omega \int_{\mathbb{R}^2} \mathcal{G}(\mathbf{y}) \left(\sum_{\mathbf{g}'\in\mathbb{Z}^2} [\gamma]_{\mathbf{g}'}\, \mathrm{e}^{i2\pi\mathbf{g}'\cdot(\mathbf{x}-\mathbf{y})} \right) d\mathbf{y}\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \\
&= \sum_{\mathbf{g}'\in\mathbb{Z}^2} [\gamma]_{\mathbf{g}'} \int_{\mathbb{R}^2} \mathcal{G}(\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{g}'\cdot\mathbf{y}}\, d\mathbf{y} \int_\Omega \mathrm{e}^{i2\pi(\mathbf{g}'-\mathbf{g})\cdot\mathbf{x}}\, d\mathbf{x} \\
&= [\gamma]_{\mathbf{g}} \int_{\mathbb{R}^2} \mathcal{G}(\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{y}}\, d\mathbf{y} \\
&= \frac{[\gamma]_{\mathbf{g}}}{2\pi\Delta^2} \int_{\mathbb{R}^2} \exp\left(-\tfrac{|\mathbf{y}|^2}{2\Delta^2} - i2\pi\mathbf{g}\cdot\mathbf{y} \right) d\mathbf{y} \\
&= \frac{[\gamma]_{\mathbf{g}}}{2\pi\Delta^2} \left(\int_{\mathbb{R}} \exp\left(-\tfrac{y_1^2}{2\Delta^2} - i2\pi g_1 y_1 \right) dy_1 \right) \left(\int_{\mathbb{R}} \exp\left(-\tfrac{y_2^2}{2\Delta^2} - i2\pi g_2 y_2 \right) dy_2 \right) \\
&= [\gamma]_{\mathbf{g}}\, \mathrm{e}^{-2\pi^2 g_1^2\Delta^2}\, \mathrm{e}^{-2\pi^2 g_2^2\Delta^2} \qquad \text{by (4.21)} \\
&= [\gamma]_{\mathbf{g}}\, \mathrm{e}^{-2\pi^2|\mathbf{g}|^2\Delta^2}.
\end{aligned}$$

Part 2. Recall the definition of H_p^s in Definition 3.23 (includes definition of $|\cdot|_\star$).

$$\begin{aligned}
\|\gamma - \widetilde{\gamma}\|_{H_p^s}^2 &= \sum_{\mathbf{g}\in\mathbb{Z}^2} |\mathbf{g}|_\star^{2s} |[\gamma - \widetilde{\gamma}]_{\mathbf{g}}|^2 \\
&= \sum_{\mathbf{g}\in\mathbb{Z}^2} |\mathbf{g}|_\star^{2s} \left(1 - \mathrm{e}^{-2\pi^2\Delta^2|\mathbf{g}|^2} \right)^2 |[\gamma]_{\mathbf{g}}|^2 \\
&= \sum_{n=1}^{\infty} \sum_{|g_1|+|g_2|=n} |\mathbf{g}|^{2s} \left(1 - \mathrm{e}^{-2\pi^2\Delta^2|\mathbf{g}|^2} \right)^2 |[\gamma]_{\mathbf{g}}|^2 \\
&\lesssim \sum_{n=1}^{\infty} n^{2s} \left(1 - \mathrm{e}^{-2\pi^2\Delta^2 n^2} \right)^2 \sum_{|g_1|+|g_2|=n} |[\gamma]_{\mathbf{g}}|^2 \\
&= \sum_{n=1}^{\infty} n^{2s} \left(1 - \mathrm{e}^{-2\pi^2\Delta^2 n^2} \right)^2 C_n^2 \qquad \text{with } C_n^2 = \textstyle\sum |[\gamma]_{\mathbf{g}}|^2 \\
&\lesssim \sum_{n=1}^{\infty} n^{2s-2} \left(1 - \mathrm{e}^{-2\pi^2\Delta^2 n^2} \right)^2 \qquad \text{since } C_n = \mathcal{O}(n^{-1}) \text{ by Theorem 3.47.}
\end{aligned} \tag{4.22}$$

To bound the expression above we need to consider the function $f(t) = 1 - \mathrm{e}^{-t^2}$. By expanding e^{-t^2} in the usual way it can be shown that if $\frac{t^4}{2!} \geq \frac{t^6}{3!}$ or $|t| \leq \sqrt{3}$ then

$$f(t) = t^2 - \tfrac{t^4}{2!} + \tfrac{t^6}{3!} - \tfrac{t^8}{4!} + \tfrac{t^{10}}{5!} - \cdots = t^2 - \left(\tfrac{t^4}{2!} - \tfrac{t^6}{3!} \right) - \left(\tfrac{t^8}{4!} - \tfrac{t^{10}}{5!} \right) - \cdots \leq t^2$$

Therefore,

$$1-\mathrm{e}^{-2\pi^2\Delta^2x^2}=f(\sqrt{2}\pi\Delta x)\leq\begin{cases}2\pi^2\Delta^2x^2 & \text{if } x^2\leq\frac{3}{2\pi^2\Delta^2}\\ 1 & \text{for all } x\in\mathbb{R}\end{cases}. \tag{4.23}$$

From (4.22) and (4.23) it follows that

$$\begin{aligned}\|\gamma-\widetilde{\gamma}\|_{H_p^s}^2&\lesssim\sum_{n=1}^{\infty}n^{2s-2}f(\sqrt{2}\pi\Delta n)^2\\&\leq 4\pi^4\Delta^4\underbrace{\sum_{n=1}^{\lfloor\frac{1}{\pi\Delta}\rfloor}n^{2s+2}}_{I_1}+\underbrace{\sum_{n=\lceil\frac{1}{\pi\Delta}\rceil}^{\infty}n^{2s-2}}_{I_2}.\end{aligned} \tag{4.24}$$

We now consider I_1 and I_2 seprately. First, consider I_1 for $-1\leq s<1/2$,

$$\begin{aligned}I_1&=4\pi^4\Delta^4\sum_{n=1}^{\lfloor\frac{1}{\pi\Delta}\rfloor}n^{2s+2}\\&=4\pi^4\Delta^4\sum_{n=1}^{\lfloor\frac{1}{\pi\Delta}\rfloor-1}n^{2s+2}+4\pi^4\Delta^4\left\lfloor\tfrac{1}{\pi\Delta}\right\rfloor^{2s+2}\\&\leq 4\pi^4\Delta^4\int_1^{\frac{1}{\pi\Delta}}x^{2s+2}dx+4\pi^4\Delta^4\left(\tfrac{1}{\pi\Delta}\right)^{2s+2}&&\text{by Lemma 3.9}\\&\leq\tfrac{4\pi^4\Delta^4}{2s+3}\left((\pi\Delta)^{-2s-3}-1\right)+4(\pi\Delta)^{2-2s}\\&=\tfrac{4(\pi\Delta)^{1-2s}}{2s+3}-\tfrac{4\pi^4\Delta^4}{2s+3}+4(\pi\Delta)^{2-2s}\\&\lesssim\Delta^{1-2s}.\end{aligned}$$

Now consider I_1 for $-3/2<s<-1$.

$$\begin{aligned}I_1&=4\pi^4\Delta^4\sum_{n=1}^{\lfloor\frac{1}{\pi\Delta}\rfloor}n^{2s+2}\\&=4\pi^4\Delta^4+\sum_{n=2}^{\lfloor\frac{1}{\pi\Delta}\rfloor}n^{2s+2}\\&\leq 4\pi^4\Delta^4+4\pi^4\Delta^4\int_1^{\frac{1}{\pi\Delta}}x^{2s+2}dx&&\text{by Lemma 3.9}\\&=4\pi^4\Delta^4+\tfrac{4\pi^4\Delta^4}{2s+3}\left((\pi\Delta)^{-2s-3}-1\right)dx\\&=4\pi^4\Delta^4+\tfrac{4(\pi\Delta)^{1-2s}}{2s+3}-\tfrac{4\pi^4\Delta^4}{2s+3}\\&\lesssim\Delta^{1-2s}.\end{aligned}$$

Therefore,

$$I_1 \lesssim \Delta^{1-2s} \qquad \text{for all } -3/2 < s < 1/2. \tag{4.25}$$

Now consider I_2. For $-3/2 < s < 1/2$ we get

$$\begin{aligned} I_2 &= \sum_{n=\lceil \frac{1}{\pi\Delta} \rceil}^{\infty} n^{2s-2} \\ &\le \left\lceil \tfrac{1}{\pi\Delta} \right\rceil^{2s-2} + \int_{\lceil \frac{1}{\pi\Delta} \rceil}^{\infty} x^{2s-2} dx \qquad \text{by Lemma 3.9} \\ &\le \left(\frac{1}{\pi\Delta}\right)^{2s-2} + \int_{\frac{1}{\pi\Delta}}^{\infty} x^{2s-2} dx \\ &= (\pi\Delta)^{2-2s} + \frac{1}{2s-1}\left(0 - (\pi\Delta)^{1-2s}\right) \\ &\lesssim \Delta^{1-2s} \end{aligned} \tag{4.26}$$

Putting (4.24), (4.25) and (4.26) together we get

$$\|\gamma - \widetilde{\gamma}\|_{H_p^s}^2 \lesssim I_1 + I_2 \lesssim \Delta^{1-2s} \qquad \text{for } -\tfrac{3}{2} < s < \tfrac{1}{2}.$$

The result then follows by taking the square root of both sides.
Part 3. For $s > 1/2$ we get

$$\begin{aligned} \|\widetilde{\gamma}\|_{H_p^s}^2 &= \sum_{\mathbf{g}\in\mathbb{Z}^2} |\mathbf{g}|_\star^{2s} |[\widetilde{\gamma}]_{\mathbf{g}}|^2 \\ &= \sum_{\mathbf{g}\in\mathbb{Z}^2} |\mathbf{g}|_\star^{2s}\, \mathrm{e}^{-4\pi^2\Delta^2|\mathbf{g}|^2}\, |[\gamma]_{\mathbf{g}}|^2 \qquad \text{by Part 1} \\ &\le |[\gamma]_0|^2 + \sum_{n=1}^{\infty} \sum_{|g_1|+|g_2|=n} |\mathbf{g}|^{2s}\, \mathrm{e}^{-2\pi^2\Delta^2|\mathbf{g}|^2}\, |[\gamma]_{\mathbf{g}}|^2 \\ &\le |[\gamma]_0|^2 + \sum_{n=1}^{\infty} n^{2s}\, \mathrm{e}^{-2\pi^2\Delta^2 n^2}\, C_n^2 \qquad \text{with } C_n^2 = \textstyle\sum |[\gamma]_{\mathbf{g}}|^2 \\ &\lesssim 1 + \sum_{n=1}^{\infty} n^{2s-2}\, \mathrm{e}^{-2\pi^2\Delta^2 n^2} \qquad \text{since } C_n = \mathcal{O}(|n|^{-1}) \text{ by Theorem 3.47.} \end{aligned} \tag{4.27}$$

Now we must consider the cases $1/2 < s \le 1$ and $s > 1$ separately. Let $f(t) = t^{2s-2}\, \mathrm{e}^{-2\pi^2\Delta^2 t^2}$. If $1/2 < s \le 1$ then $f(t)$ is monotonically decreasing for $t > 0$ and using Lemma 3.9 we get

$$\sum_{n=1}^{\infty} n^{2s-2}\, \mathrm{e}^{-2\pi^2\Delta^2 n^2} \le \int_0^{\infty} x^{2s-2}\, \mathrm{e}^{-2\pi^2\Delta^2 x^2}\, dx \tag{4.28}$$

Alternatively, if $s > 1$ then $f(t)$ (for $t \ge 0$) has a single maximum at $t_0 = \frac{\sqrt{2s-2}}{2\pi\Delta}$,

and is monotonically increasing on the interval $[0, t_0]$ and monotonically decreasing on $[t_0, \infty)$. Moreover, $f(t_0) \lesssim \Delta^{2-2s}$. Therefore, Lemma 3.9 gives us

$$\begin{aligned}\sum_{n=1}^{\infty} n^{2s-2}\,\mathrm{e}^{-2\pi^2\Delta^2 n^2} &= \sum_{n=1}^{\lfloor t_0\rfloor-1} f(n) + f(\lfloor t_0\rfloor) + f(\lceil t_0\rceil) + \sum_{\lceil t_0\rceil+1} f(n)\\ &\le \int_1^{\lfloor t_0\rfloor} f(x)dx + 2f(t_0) + \int_{\lceil t_0\rceil}^{\infty} f(x)dx\\ &\lesssim \Delta^{2-2s} + \int_0^{\infty} x^{2s-2}\,\mathrm{e}^{-2\pi^2\Delta^2x^2}\,dx \qquad (4.29)\end{aligned}$$

Now put (4.27), (4.28) and (4.29) together to get, for $s > 1/2$,

$$\begin{aligned}\|\widetilde{\gamma}\|_{H_p^s}^2 &\lesssim 1 + \Delta^{2-2s} + \int_0^{\infty} x^{2s-2}\,\mathrm{e}^{-2\pi^2\Delta^2x^2}\,dx\\ &= 1 + \Delta^{2-2s} + \frac{1}{\Delta^{2s-1}}\int_0^{\infty} y^{2s-2}\,\mathrm{e}^{-2\pi^2y^2}\,dy \qquad \text{substituting } y = \Delta x\\ &\lesssim \Delta^{1-2s} \qquad \text{since the integral is bounded independent of } \Delta.\end{aligned}$$

Therefore, $\|\widetilde{\gamma}\|_{H_p^s} \lesssim \Delta^{-s+1/2}$ for $s > 1/2$. Now consider the case when $s = 1/2$. Following the same argument to that in (4.27) we get

$$\begin{aligned}\|\widetilde{\gamma}\|_{H_p^{1/2}}^2 &\lesssim 1 + \sum_{n=1}^{\infty} n^{-1}\,\mathrm{e}^{-2\pi^2\Delta^2n^2} \le 2 + \sum_{n=2}^{\infty} n^{-1}\,\mathrm{e}^{-2\pi^2\Delta^2n^2}\\ &\le 2 + \int_1^{\infty} x^{-1}\,\mathrm{e}^{-2\pi^2\Delta^2x^2}\,dx \qquad \text{by Lemma 3.9}\\ &= 2 + \int_{\Delta}^{\infty} y^{-1}\,\mathrm{e}^{-2\pi^2y^2}\,dy \qquad \text{substituting } y = \Delta x\\ &= 2 + \int_{\Delta}^{1} y^{-1}\,\mathrm{e}^{-2\pi^2y^2}\,dy + \int_1^{\infty} y^{-1}\,\mathrm{e}^{-2\pi^2y^2}\,dy\\ &\le 2 + \int_{\Delta}^{1} y^{-1}dy + \int_1^{\infty} y^{-1}\,\mathrm{e}^{-2\pi^2y^2}\,dy\\ &= 2 + \log(\Delta^{-1}) + \int_1^{\infty} y^{-1}\,\mathrm{e}^{-2\pi^2y^2}\,dy\\ &\lesssim \log(\Delta^{-1}).\end{aligned}$$

Therefore, $\|\widetilde{\gamma}\|_{H_p^{1/2}} \lesssim \sqrt{\log(\Delta^{-1})}$.

Finally, for $s < 1/2$ we get

$$\begin{aligned}\|\widetilde{\gamma}\|_{H_p^s}^2 &= \sum_{\mathbf{g}\in\mathbb{Z}^2} |\mathbf{g}|_\star^{2s}|[\widetilde{\gamma}]_{\mathbf{g}}|^2 = \sum_{\mathbf{g}\in\mathbb{Z}^2} |\mathbf{g}|_\star^{2s}\,\mathrm{e}^{-4\pi^2\Delta^2|\mathbf{g}|^2}\,|[\gamma]_{\mathbf{g}}|^2 \qquad \text{by Part 1}\\ &\le \sum_{\mathbf{g}\in\mathbb{Z}^2} |\mathbf{g}|_\star^{2s}|[\gamma]_{\mathbf{g}}|^2 = \|\gamma\|_{H_p^s}^2.\end{aligned}$$

Therefore, $\|\widetilde{\gamma}\|_{H_p^s} \leq \|\gamma\|_{H_p^s}$ for $s < 1/2$. Since $\gamma \in H_p^s$ for $s < 1/2$ by Theorem 3.40, we get $\|\widetilde{\gamma}\|_{H_p^s} \lesssim 1$ for $s < 1/2$. □

The results from Lemma 4.26 have analogous results in 1D and the proofs use the same techniques.

We now define the smooth problem. The operator L is modified and we define the modified opeator $\widetilde{\mathrm{L}}$ as

$$\widetilde{\mathrm{L}} = -\nabla^2 - \widetilde{\gamma}(\mathbf{x}) + \mathrm{K}$$

which is the same as the operator in (4.2) except $\gamma(\mathbf{x})$ has been replaced with $\widetilde{\gamma}(\mathbf{x})$. As in the previous section we consider this operator on the Hilbert space $L^2(\mathbb{R}^2)$. We apply the Floquet tranform to $\widetilde{\mathrm{L}}$ in just the same way as in Subsection 4.1.2 and it is possible to show that all of the results from Subsection 4.1.2 that were given for L also apply for $\widetilde{\mathrm{L}}$ and the proofs are the same. Just as in Subsection 4.1.3 for L and $\mathrm{L}_{\boldsymbol{\xi}}$ we define the variational form of the smooth problem as

Problem 4.27. For a fixed $\boldsymbol{\xi} \in B$, find an eigenpair $(\tilde{\lambda}, u)$ where $\tilde{\lambda} \in \mathbb{C}$ and $0 \neq u \in H_p^1$ such that

$$\tilde{a}(u, v) = \tilde{\lambda} b(u, v) \qquad \forall v \in H_p^1 \tag{4.30}$$

where

$$\tilde{a}(u, v) = \int_\Omega (\nabla + i\xi)\, u \cdot \overline{(\nabla + i\xi)\, v} + (\mathrm{K} - \widetilde{\gamma})\, u\overline{v} dx$$

and $b(\cdot, \cdot)$ is the same as in Problem 4.6.

The method is to now approximate the solution to Problem 4.27 via the spectral Galerkin method of Section 4.2. We replace H_p^1 with $\mathcal{S}_G$ in Problem 4.27 to get the corresponding discrete variational eigenvalue problem,

Problem 4.28. Find $\tilde{\lambda}_G \in \mathbb{R}$ and $0 \neq u_G \in \mathcal{S}_G$ such that

$$\tilde{a}(u_G, v_G) = \tilde{\lambda}_G b(u_G, v_G) \qquad \forall v_G \in \mathcal{S}_G. \tag{4.31}$$

As in Section 4.2 we can write this problem as a matrix eigenvalue problem and we solve it using the same implementation as we did for the original problem.

Using the same proof techniques as in Theorem 4.22 and Theorem 4.21 we can show that exactly the same preconditioning results hold. We now develop the error analysis to include smoothing in the next section.

4.3.2 Error Analysis

In this subsection we bound the error between the eigenvalues and eigenfunctions of Problem 4.6 and Problem 4.28. To do this we consider Problem 4.27 as an intermediate problem and we express the error between Problem 4.6 and Problem 4.28 as the sum

of two separate error contributions. The first contribution is the smoothing error that was introduced when we replaced piecewise constant $\gamma(\mathbf{x})$ with a smooth function $\widetilde{\gamma}(\mathbf{x})$. This is measured by considering the difference in the solutions of Problem 4.6 and Problem 4.27. The second error contribution comes from our spectral Galerkin method. This is measured by considering the difference between the solutions of Problem 4.27 and Problem 4.28.

Before we prove any error bounds we must first prove the following lemma.

Lemma 4.29. *Problem 4.27 (with $\gamma \in PC'_p$) has the following properties:*

1. *The bilinear form $\tilde{a}(\cdot,\cdot)$ is bounded, coercive and Hermitian.*
2. *The bilinear form $\tilde{a}(\cdot,\cdot)$ defines an inner product on H^1_p which has an induced norm $\|\cdot\|_{\tilde{a}} := |\tilde{a}(\cdot,\cdot)|^{1/2}$ that is equivalent to $\|\cdot\|_{H^1_p}$.*
3. *The solution operator corresponding to Problem 4.27, $\widetilde{\mathrm{T}} : H^1_p \to H^1_p$, is bounded, positive, compact and self-adjoint with respect to $\tilde{a}(\cdot,\cdot)$.*
4. *Problem 4.27 has a countable set of real eigenvalues that are positive and the corresponding eigenfunctions can be chosen so that they are orthogonal with respect to $\tilde{a}(\cdot,\cdot)$ and they are complete in L^2_p.*
5. *If u is an eigenfunction of Problem 4.27 (with $\gamma \in PC'_p$) then $u \in C^\infty_p$ and*

$$\|u\|_{H^s_p} \lesssim \begin{cases} \|u\|_{H^1_p} & \text{for } s < \frac{5}{2} \\ \sqrt{\log(\Delta^{-1})}\|u\|_{H^1_p} & \text{for } s = \frac{5}{2} \\ \Delta^{-s+5/2}\|u\|_{H^1_p} & \text{for } s > \frac{5}{2} \end{cases}$$

Proof. We only prove Part 5 as the proofs for Parts 1-4 are the same as the proofs for Lemmas 4.7, 4.9 and 4.10.

Let $\tilde{\lambda}$ be the eigenvalue of Problem 4.27 that corresponds to the eigenfunction u. Since u is an eigenfunction of Problem 4.27 we have that u is a weak solution of an elliptic boundary value problem of the same form as (3.52) with $L := \widetilde{\mathrm{L}}_{\boldsymbol{\xi}}$ and $f := \tilde{\lambda}u$ where L is elliptic with C^∞_p coefficients. Using Theorem 3.77 we can "boot-strap" our way to $u \in H^s_p$ for any $s \in \mathbb{R}$. We then use Theorem 3.27 to get $u \in C^\infty_p$.

To obtain the estimates of $\|u\|_{H^s_p}$ in Part 5 of our lemma we consider a new boundary value problem of the same form as (3.52). Now let $L := -(\nabla + i\boldsymbol{\xi})^2 + \mathrm{K}$ and $f := \tilde{\lambda}u + \widetilde{\gamma}u$. Again L is elliptic, and now it has constant coefficients. u is a weak solution to this boundary value problem.

First, let us bound $\|f\|_{L^2_p}$. $\|f\|_{L^2_p} \leq |\lambda|\|u\|_{L^2_p} + \|\widetilde{\gamma}\|_\infty \|u\|_{L^2_p} \lesssim \|u\|_{H^1_p}$ since $\widetilde{\gamma}$ is continuous. Theorem 3.77 implies that

$$\|u\|_{H^2_p} \lesssim \|u\|_{H^1_p}. \tag{4.32}$$

Now consider $\|f\|_{H_p^s}$ for $s < \frac{1}{2}$. We have

$$\begin{aligned} \|f\|_{H_p^s} &\lesssim \|u\|_{H_p^s} + \|\widetilde{\gamma}\|_{H_p^s}\|u\|_{H_p^2} && \text{by Theorem 3.28} \\ &\lesssim \|u\|_{H_p^1} && \text{by Lemma 4.26 and (4.32).} \end{aligned}$$

Theorem 3.77 now implies that

$$\|u\|_{H_p^s} \lesssim \|u\|_{H_p^1} \qquad \text{for } s < \tfrac{5}{2}. \tag{4.33}$$

Now consider $\|f\|_{H_p^s}$ for $\frac{1}{2} \le s < \frac{5}{2}$. We have

$$\begin{aligned} \|f\|_{H_p^s} &\lesssim \begin{cases} \|u\|_{H_p^s} + \|\widetilde{\gamma}\|_{H_p^s}\|u\|_{H_p^2} & \frac{1}{2} \le s \le 1 \\ \|u\|_{H_p^s} + \|\widetilde{\gamma}\|_{H_p^s}\|u\|_{H_p^s} & 1 < s < \frac{5}{2} \end{cases} && \text{by Theorem 3.28} \\ &\lesssim \begin{cases} \sqrt{\log(\Delta^{-1})}\|u\|_{H_p^1} & s = \frac{1}{2} \\ \Delta^{-s+1/2}\|u\|_{H_p^1} & \frac{1}{2} < s < \frac{5}{2} \end{cases} && \text{by (4.33) and Lemma 4.26.} \end{aligned}$$

We apply Theorem 3.77 once again to get

$$\|u\|_{H_p^s} \lesssim \begin{cases} \sqrt{\log(\Delta^{-1})}\|u\|_{H_p^1} & s = \frac{5}{2} \\ \Delta^{-s+5/2}\|u\|_{H_p^1} & \frac{5}{2} < s < \frac{9}{2}. \end{cases} \tag{4.34}$$

We now use induction to prove that $\|u\|_{H_p^s} \lesssim \Delta^{-s+5/2}\|u\|_{H_p^1}$ for $s \in \mathbb{N}$, $s \ge 4$. We have already proved the $s = 4$ case in (4.34). Our inductive hypothesis is to assume that for $k \in \mathbb{N}$,

$$\|u\|_{H_p^s} \lesssim \Delta^{-s+5/2}\|u\|_{H_p^1} \qquad \text{for } s \in \mathbb{N},\ 4 \le s \le k. \tag{4.35}$$

Consider $\|f\|_{H_p^{k-1}}$. By (4.35) we get

$$\|f\|_{H_p^{k-1}} \lesssim \|u\|_{H_p^{k-1}} + \|\widetilde{\gamma}u\|_{H_p^{k-1}} \lesssim \Delta^{-k+3/2}\|u\|_{H_p^1} + \|\widetilde{\gamma}u\|_{H_p^{k-1}} \tag{4.36}$$

The key is to now bound $\|\widetilde{\gamma}u\|_{H_p^{k-1}}$ in an efficient way. We do not use Theorem 3.28 because the bound is not sharp enough. Instead we do the following. Let α and β

define multi-indices. We write $\alpha \leq \beta$, for $\alpha_i \leq \beta_i$ for all i.

$$
\begin{aligned}
\|\widetilde{\gamma}u\|^2_{H^{k-1}_p} &= \|\widetilde{\gamma}u\|^2_{H^{k-1}(\Omega)} \\
&= \sum_{|\alpha|\leq k-1} \|D^\alpha(\widetilde{\gamma}u)\|^2_{L^2(\Omega)} \\
&= \sum_{|\alpha|\leq k-1} \left\| \sum_{\beta\leq\alpha} \binom{\alpha}{\beta} \left(D^\beta \widetilde{\gamma}\right) \left(D^{\alpha-\beta}u\right) \right\|^2_{L^2(\Omega)} \\
&\lesssim \sum_{|\alpha|\leq k-1} \sum_{\beta\leq\alpha} \|(D^\beta\widetilde{\gamma})(D^{\alpha-\beta}u)\|^2_{L^2(\Omega)} \\
&= \sum_{|\alpha|\leq k-1} \sum_{j=0}^{|\alpha|} \sum_{\substack{|\beta|=j \\ \beta\leq\alpha}} \|(D^\beta\widetilde{\gamma})(D^{\alpha-\beta}u)\|^2_{L^2(\Omega)} \\
&= \sum_{|\alpha|\leq k-1} \left(\|\widetilde{\gamma}D^\alpha u\|^2_{L^2(\Omega)} + \sum_{j=1}^{|\alpha|} \sum_{\substack{|\beta|=j \\ \beta\leq\alpha}} \|(D^\beta\widetilde{\gamma})(D^{\alpha-\beta}u)\|^2_{L^2(\Omega)} \right) \\
&\leq \sum_{|\alpha|\leq k-1} \left(\|\widetilde{\gamma}\|^2_\infty \|D^\alpha u\|^2_{L^2(\Omega)} + \sum_{j=1}^{|\alpha|} \sum_{\substack{|\beta|=j \\ \beta\leq\alpha}} \|D^\beta\widetilde{\gamma}\|^2_{L^2(\Omega)} \|D^{\alpha-\beta}u\|^2_\infty \right) \\
&\lesssim \|\widetilde{\gamma}\|^2_\infty \|u\|^2_{H^{k-1}(\Omega)} + \sum_{|\alpha|\leq k-1} \sum_{j=1}^{|\alpha|} \|\widetilde{\gamma}\|^2_{H^j(\Omega)} \max_{|\beta|\leq|\alpha|-j} \|D^\beta u\|^2_\infty \\
&\lesssim \|\widetilde{\gamma}\|^2_\infty \|u\|^2_{H^{k-1}(\Omega)} + \sum_{|\alpha|\leq k-1} \sum_{j=1}^{|\alpha|} \|\widetilde{\gamma}\|^2_{H^j(\Omega)} \max_{|\beta|\leq|\alpha|-j} \|D^\beta u\|^2_{H^2(\Omega)} \quad \text{by Thm. 3.27} \\
&\leq \|\widetilde{\gamma}\|^2_\infty \|u\|^2_{H^{k-1}(\Omega)} + \sum_{|\alpha|\leq k-1} \sum_{j=1}^{|\alpha|} \|\widetilde{\gamma}\|^2_{H^j(\Omega)} \|u\|^2_{H^{|\alpha|-j+2}(\Omega)} \\
&\lesssim \|\widetilde{\gamma}\|^2_\infty \|u\|^2_{H^{k-1}(\Omega)} + \sum_{j=1}^{k-1} \|\widetilde{\gamma}\|^2_{H^j(\Omega)} \|u\|^2_{H^{k-j+1}(\Omega)} \\
&= \|\widetilde{\gamma}\|^2_\infty \|u\|^2_{H^{k-1}(\Omega)} + \sum_{j=1}^{k-2} \|\widetilde{\gamma}\|^2_{H^j(\Omega)} \|u\|^2_{H^{k-j+1}(\Omega)} + \|\widetilde{\gamma}\|^2_{H^{k-1}(\Omega)} \|u\|^2_{H^2(\Omega)} \\
&\lesssim \left(\Delta^{-2k+3} + \sum_{j=1}^{k-2} \Delta^{-2j+1}\Delta^{-2(k-j+1)+5} + \Delta^{-2k+3} \right) \|u\|_{H^1_p} \\
&\qquad\qquad \text{by (4.35) and Lemma 4.26} \\
&= \left(\Delta^{-2k+3} + \sum_{j=1}^{k-2} \Delta^{-2k+4} + \Delta^{-2k+3} \right) \|u\|_{H^1_p} \\
&\lesssim \Delta^{-2k+3} \|u\|_{H^1_p}.
\end{aligned}
$$

Putting this back into (4.36) we get

$$\|f\|_{H_p^{k-1}} \lesssim \Delta^{-k+3/2}\|u\|_{H_p^1}.$$

Theorem 3.77 implies that

$$\|u\|_{H_p^{k+1}} \lesssim \Delta^{-k+3/2}\|u\|_{H_p^1} = \Delta^{-(k+1)+5/2}\|u\|_{H_p^1}.$$

Therefore, by induction, (4.33) and (4.34) we have

$$\|u\|_{H_p^s} \lesssim \begin{cases} \|u\|_{H_p^s} & s < \frac{5}{2} \\ \sqrt{\log(\Delta^{-1})}\|u\|_{H_p^1} & s = \frac{5}{2} \\ \Delta^{-s+5/2}\|u\|_{H_p^1} & \in \left(\frac{5}{2}, \frac{9}{2}\right) \cup \{s \in \mathbb{N} : s \geq 5\}. \end{cases}$$

The result then follows by applying Lemma 3.26. □

The first error contribution we examine is that of smoothing. We bound the difference between Problem 4.6 and Problem 4.27. These are both infinite dimensional problems but we can still apply Theorem 3.68. To do this T and $\widetilde{\mathrm{T}}$ must satisfy the conditions of Theorem 3.68. We must show that $\widetilde{\mathrm{T}} \to \mathrm{T}$ in norm as $\Delta \to 0$. This property is proved using the following lemma. The proof will use Strang's 1st Lemma (Theorem 3.75) in a non-standard way in the sense that we apply it when an infinite dimensional problem approximates another infinite dimensional problem.

Lemma 4.30. *For $\Delta \geq 0$ (and $\gamma \in PC_p'$) we get:*

1. $$\|\mathrm{T} - \widetilde{\mathrm{T}}\|_{H_p^1} \lesssim \Delta^{3/2}.$$

2. *The adjoint of $\widetilde{\mathrm{T}}$ with respect to $a(\cdot,\cdot)$, $\widetilde{\mathrm{T}}^*$, satisfies*
$$\|\mathrm{T} - \widetilde{\mathrm{T}}^*\|_{H_p^1} \lesssim \Delta^{3/2}.$$

3. *For $u, v \in H_p^1$,*
$$\left|a((\mathrm{T} - \widetilde{\mathrm{T}})u, v)\right| \lesssim \Delta^{3/2}\|u\|_{H_p^1}\|v\|_{H_p^1}.$$

Proof. Part 1. The proof for this result relies on Strang's 1st Lemma (Theorem 3.75).

Let $f \in H_p^1$. Then using Theorem 3.75 we get

$$\begin{aligned}
\left\|\mathrm{T} f - \widetilde{\mathrm{T}} f\right\|_{H_p^1} &\lesssim \inf_{v \in H_p^1} \left\{ \|\mathrm{T} f - v\|_{H_p^1} + \sup_{w \in H_p^1} \frac{|a(v,w) - \tilde{a}(v,w)|}{\|w\|_{H_p^1}} \right\} \\
&\leq \sup_{w \in H_p^1} \frac{|a(\mathrm{T} f, w) - \tilde{a}(\mathrm{T} f, w)|}{\|w\|_{H_p^1}} && \text{choosing } v = Tf \\
&\leq \sup_{w \in H_p^1} \frac{\int_\Omega |(\widetilde{\gamma} - \gamma)\, \mathrm{T} f \overline{w}| dx}{\|w\|_{H_p^1}} \\
&\leq \sup_{w \in H_p^1} \frac{\|\mathrm{T} f\|_\infty \int_\Omega |(\widetilde{\gamma} - \gamma) \overline{w}| dx}{\|w\|_{H_p^1}} \\
&\leq \sup_{w \in H_p^1} \frac{\|\mathrm{T} f\|_\infty \|\widetilde{\gamma} - \gamma\|_{H_p^{-1}} \|w\|_{H_p^1}}{\|w\|_{H_p^1}} && \text{by (3.3)} \\
&= \|\mathrm{T} f\|_\infty \|\widetilde{\gamma} - \gamma\|_{H_p^{-1}} \\
&\lesssim \|\mathrm{T} f\|_{H_p^2} \|\widetilde{\gamma} - \gamma\|_{H_p^{-1}} && \text{by Theorem 3.27} \\
&\lesssim \|f\|_{H_p^1} \|\widetilde{\gamma} - \gamma\|_{H_p^{-1}} && \text{by Theorem 4.11} \\
&\lesssim \|f\|_{H_p^1} \Delta^{3/2} && \text{by Lemma 4.26.}
\end{aligned}$$

Part 2. The proof of Part 2 uses Part 1 and the fact that $a(\cdot,\cdot)$ is bounded in H_p^1. For $f \in H_p^1$ we get

$$\begin{aligned}
\|(\mathrm{T} - \widetilde{\mathrm{T}}^*) f\|_{H_p^1}^2 &\lesssim \|(\mathrm{T} - \widetilde{\mathrm{T}}^*) f\|_a^2 \\
&= a((\mathrm{T} - \widetilde{\mathrm{T}}^*) f, (\mathrm{T} - \widetilde{\mathrm{T}}^*) f) \\
&= a((\mathrm{T} - \widetilde{\mathrm{T}})(\mathrm{T} - \widetilde{\mathrm{T}}^*) f, f) \\
&\lesssim \|(\mathrm{T} - \widetilde{\mathrm{T}})(\mathrm{T} - \widetilde{\mathrm{T}}^*) f\|_{H_p^1} \|f\|_{H_p^1} \\
&\leq \|\mathrm{T} - \widetilde{\mathrm{T}}\|_{H_p^1} \|(\mathrm{T} - \widetilde{\mathrm{T}}^*) f\|_{H_p^1} \|f\|_{H_p^1}.
\end{aligned}$$

By dividing through by $\|(\mathrm{T} - \widetilde{\mathrm{T}}^*) f\|_{H_p^1}$ we get

$$\|(\mathrm{T} - \widetilde{\mathrm{T}}^*) f\|_{H_p^1} \lesssim \|\mathrm{T} - \widetilde{\mathrm{T}}\|_{H_p^1} \|f\|_{H_p^1}.$$

The result then follows by using Part 1.

Part 3. The proof of Part 3 follows directly from Part 1 using the fact that $a(\cdot,\cdot)$ is bounded. □

We now apply Theorem 3.68 to obtain bounds on the eigenvalue and eigenfunction errors of Problem 4.27 as an approximation of Problem 4.6 for sufficiently small Δ.

Theorem 4.31. *Let λ be an eigenvalue of Problem 4.6 (with $\gamma \in PC_p'$) with multiplicity m and corresponding eigenspace M. Then for sufficiently small Δ there exist m*

eigenvalues $\tilde{\lambda}_1(\Delta), \ldots, \tilde{\lambda}_m(\Delta)$ *(counted according to multiplicity) of Problem 4.27 with corresponding eigenspaces* $M_1(\tilde{\lambda}_1), \ldots, M_m(\tilde{\lambda}_m)$ *and a space*

$$\mathcal{M}_\Delta := \bigoplus_{j=1}^{m} M_j(\tilde{\lambda}_j)$$

such that

$$\delta(M, \mathcal{M}_\Delta) \lesssim \Delta^{3/2}$$

and

$$|\lambda - \tilde{\lambda}_j| \lesssim \Delta^{3/2} \qquad \text{for } j = 1, \ldots, m.$$

Proof. The proof of this result is very similar to the proof of Theorem 4.24. First we check that the conditions of Theorem 3.68 are satisfied. Just as in the proof of Theorem 4.24, H_p^1 is our Hilbert space with inner product $a(\cdot,\cdot)$. T is bounded, compact and self-adjoint. $\widetilde{\mathrm{T}}$ ($\Delta > 0$) is a family of bounded compact operators (Lemma 4.29) and Part 1 of Lemma 4.30 ensures that $\widetilde{\mathrm{T}} \to \mathrm{T}$ in norm as $\Delta \to 0$. $\widetilde{\mathrm{T}}$ is not self-adjoint with respect to $a(\cdot,\cdot)$ but it is self-adjoint with respect to $\tilde{a}(\cdot,\cdot)$. $\widetilde{\mathrm{T}}$ bounded, compact and self-adjoint (with respect to $\tilde{a}(\cdot,\cdot)$) ensures that $\widetilde{\mathrm{T}}$ does not have any generalised eigenvectors. Now we apply Theorem 3.68, Lemma 3.71 and Lemma 4.30 to obtain the result. □

We now have a result that quantifies the difference between Problem 4.6 and 4.27. As we expect, as $\Delta \to 0$ the eigenvalues and eigenfunctions of the smooth problem converge to the eigenvalues and eigenfunctions of our original problem. However, we might have expected to obtain an eigenvalue estimate that decreased at twice the rate of the eigenfunction error, as we did in Theorem 4.24. We have not been able to prove this type of result because there is no "Galerkin orthogonality" condition that the eigenfunctions of both problems satisfy. Later, numerical results will show that the eigenvalue errors do not decrease at twice the rate of the eigenfunction errors. However, the numerical results will show that our result is not completely sharp for the eigenvalue error estimate. Theorem 4.31 also holds for the 1D problem.

We are now free to concentrate on the error that we introduce when we approximate Problem 4.27 with a discrete problem, Problem 4.28. We studied this error in the previous section when we applied the spectral Galerkin method to our original problem. The error analysis for the spectral Galerkin method applied to the smooth problem is the same except for the approximation error estimate, which depends on the regularity of the eigenfunctions. We have already shown, in Lemma 4.29, that because $\widetilde{\gamma}$ is smooth, the eigenfunctions of Problem 4.27 are in C_p^∞. Therefore, we now expect the approximation error to decrease superalgebraically with respect to G (i.e. decrease with arbitrary algebraic order). However, we also expect the approximation error to depend on the amount of smoothing, Δ. We expect to see the approximation error

increase as $\Delta \to 0$ since the derivatives of the coefficient function $\widetilde{\gamma}(\mathbf{x})$ will become larger as $\Delta \to 0$. Indeed, our task will be to derive an approximation error bound that shows the dependence on G *and* Δ, which we do the following lemma. We have already done the hard work when we proved the estimates of $\|u\|_{H_p^s}$ in Part 5 of Lemma 4.29 and the following approximation error result follows neatly from this.

Lemma 4.32. *Let u be an eigenfunction of Problem 4.27 (with $\gamma \in PC_p'$). Then we obtain the following family of bounds for the approximation error,*

$$\inf_{\chi \in \mathcal{S}_G} \|u - \chi\|_{H_p^1} \lesssim \begin{cases} G^{-3/2+\epsilon}\|u\|_{H_p^1} & \text{for } \epsilon > 0 \\ G^{-3/2}\sqrt{\log(\Delta^{-1})}\|u\|_{H_p^1} & \\ G^{-3/2-s}\Delta^{-s}\|u\|_{H_p^1} & \text{for } s > 0. \end{cases}$$

Proof. This result follows from Part 5 of Lemma 4.29 and Lemma 3.30 by taking $\chi = \mathrm{P}_G^{(\mathcal{S})} u$. □

We have shown that the approximation error for eigenfunctions of Problem 4.27 and our finite dimensional space $\mathcal{S}_G$ decreases at a superalgebraic rate (arbitrary polynomial order) with respect to G. However, the fast convergence with respect to G does not come without a penalty when Δ is small. Indeed, when we take s larger in Lemma 4.32 (to obtain faster convergence with respect to G), the penalty for small Δ also becomes larger.

We now state a result for the errors of the spectral Galerkin method applied to Problem 4.27 that is similar to Theorem 4.24, except we use our new approximation error result (Lemma 4.32) to obtain different error estimates. The proof is analogous to the proof of Theorem 4.24, except we use Lemma 4.32 instead of Part 5 of Lemma 4.23.

Theorem 4.33. *Let $\tilde{\lambda}$ be an eigenvalue of Problem 4.27 (with $\gamma \in PC_p'$) with multiplicity m and corresponding eigenspace $\widetilde{M}$. Then, for sufficiently large G, there exist m eigenvalues $\tilde{\lambda}_1(G,\Delta), \ldots, \tilde{\lambda}_m(G,\Delta)$, counted according to multiplicity, of Problem 4.28 with corresponding eigenspaces $\widetilde{M}_1(\tilde{\lambda}_1), \ldots, \widetilde{M}_m(\tilde{\lambda}_m)$ and a space*

$$\widetilde{\mathcal{M}}_{G,\Delta} := \bigoplus_{j=1}^{m} \widetilde{M}_j(\tilde{\lambda}_j)$$

such that

$$\delta(\widetilde{M}, \widetilde{\mathcal{M}}_{G,\Delta}) \lesssim \begin{cases} G^{-3/2+\epsilon} & \text{for } \epsilon > 0 \\ G^{-3/2}\sqrt{\log(\Delta^{-1})} & \\ G^{-3/2-s}\Delta^{-s} & \text{for } s > 0 \end{cases}$$

and

$$|\tilde{\lambda} - \tilde{\lambda}_j| \lesssim \begin{cases} G^{-3+2\epsilon} & \text{for } \epsilon > 0 \\ G^{-3}\log(\Delta^{-1}) & \\ G^{-3-2s}\Delta^{-2s} & \text{for } s > 0 \end{cases}$$

for $j = 1, \ldots, m$.

In Theorem 4.33 we have proved that the eigenvalues and eigenfunctions of the discrete smooth problem converge superalgebraically to the eigenvalues and eigenfunctions of the exact smooth problem. Notice also that the eigenvalues converge at twice the rate of the eigenfunctions in this case.

So far we have analysed the error from modifying the original problem and we have analysed the error from solving the modified problem with the spectral Galerkin method. The next step of the smooth problem error analysis is to add the two error contributions together. We do this and get the following Theorem. The proof is omitted because it is a simple application of the triangle inequality to the results of Theorem 4.31 and Theorem 4.33.

Theorem 4.34. *Let* λ *be an eigenvalue of Problem 4.6 (with* $\gamma \in PC_p'$*) with multiplicity* m *and corresponding eigenspace* M*. Then, for sufficiently large* G *and small* $\Delta > 0$*, there exist* m *eigenvalues* $\tilde{\lambda}_1(G, \Delta), \ldots, \tilde{\lambda}_m(G, \Delta)$ *of Problem 4.28 with corresponding eigenspaces* $\widetilde{M}_1(\tilde{\lambda}_1), \ldots, \widetilde{M}_m(\tilde{\lambda}_m)$ *and a space*

$$\widetilde{\mathcal{M}}_{G,\Delta} := \bigoplus_{j=1}^{m} \widetilde{M}_j(\tilde{\lambda}_j)$$

such that

$$\delta(M, \widetilde{\mathcal{M}}_{G,\Delta}) \lesssim \begin{cases} \Delta^{3/2} + G^{-3/2+\epsilon} & \text{for } \epsilon > 0 \\ \Delta^{3/2} + G^{-3/2}\sqrt{\log(\Delta^{-1})} & \\ \Delta^{3/2} + G^{-3/2-s}\Delta^{-s} & \text{for } s > 0 \end{cases} \tag{4.37}$$

and

$$|\tilde{\lambda} - \tilde{\lambda}_j| \lesssim \begin{cases} \Delta^{3/2} + G^{-3+2\epsilon} & \text{for } \epsilon > 0 \\ \Delta^{3/2} + G^{-3}\log(\Delta^{-1}) & \\ \Delta^{3/2} + G^{-3-2s}\Delta^{-2s} & \text{for } s > 0 \end{cases} \tag{4.38}$$

for $j = 1, \ldots, m$.

The final step of the error analysis for the smoothing method is to suggest a smoothing technique based on our theoretical error bounds. We want to choose $\Delta = f(G)$ to minimise the error. As we will see, to obtain optimal error convergence rates for our method it will be sufficient to choose $\Delta = CG^r$ for some degree $r \in \mathbb{R}$ and constant C.

It is possible to approach the problem of choosing an optimal amount of smoothing from two directions. The first approach is to minimise the error bounds in Theorem 4.34 by balancing the two terms on the right-hand-sides of (4.37) and (4.38). This approach will give a value of r that produces an optimal error bound. The second approach is to remember that this method is supposed to improve the standard method (with no smoothing). With this in mind we aim to choose r so that the two error bounds in Theorem 4.34 are smaller than the corresponding error bounds from Theorem 4.24.

Corollary 4.35. *To optimize the error bounds in Theorem 4.34 with $\Delta := G^r$ we must choose*

1. *$r = -1$ to optimize the error bound for the eigenfunction errors. This gives us an error bound of*
$$\delta(M, \widetilde{\mathcal{M}}_{G,\Delta}) \lesssim G^{-3/2}$$

2. *$r = -2$ to optimize the error bound for the eigenvalue errors. This gives us an error bound of*
$$|\tilde{\lambda} - \tilde{\lambda}_j| \lesssim G^{-3+2\epsilon} \qquad \text{for } \epsilon > 0$$
and $j = 1, \ldots, m$.

Therefore, no choice of smoothing will result in an error bound that decreases at a faster rate than the error bounds for the standard method in Theorem 4.24.

Proof. We will first consider the eigenfunction error bound from Theorem 4.34. We must use the third case of (4.37) with the form

$$\Delta^{3/2} + G^{-3/2-s}\Delta^{-s} \qquad \text{for } s > 0 \tag{4.39}$$

since the first two cases of (4.37) will result in an error bound that converges slower than $\mathcal{O}(G^{-3/2})$ (which is the rate of decay of the error bound for the standard method in Theorem 4.24). We substitute $\Delta = G^r$ into (4.39) and balance the terms by equating the degree of each term. We get $\frac{3r}{2} = \frac{3}{2} - s - sr$. Solving for r we get $r = -1$ and the result follows.

We now consider the eigenvalue error bound from Theorem 4.34. We must use the third case of (4.38) where the error has the form

$$\Delta^{3/2} + G^{-3-2s}\Delta^{-2s} \qquad \text{for } s > 0 \tag{4.40}$$

since the first two cases of (4.38) cannot give us an error bound that converges faster than $\mathcal{O}(G^{-3})$ which is the rate of decay of the error bound for the standard method in Theorem 4.24). We substitute $\Delta = G^r$ into (4.40) and balance the terms by equating

the degree of each term. We get $\frac{3r}{2} = -3 - 2s - 2sr$. Solving for r we get

$$r = -\left(1 + \frac{3}{3+4s}\right). \tag{4.41}$$

With this choice of r for $\Delta = G^r$ we get eigenvalue errors that have $\mathcal{O}(G^{-\frac{3}{2}(1+\frac{3}{3+4s})})$ for $s > 0$. Choosing $s \to 0$ we get the fastest rate of decay and the eigenvalues errors decrease with a rate that approaches $\mathcal{O}(G^{-3})$.

In fact, if we choose $r = -2$ then we get eigenvalue error of $\mathcal{O}(G^{-3+2s})$ which also approaches $\mathcal{O}(G^{-3})$ as $s \to 0$ and is also optimal. □

The previous corollary contains the main conclusion of this section, "No choice of smoothing will give us an error bound that decays faster than the error bound for the standard method". It also gives specific values of r in $\Delta = G^r$ that will recover the decay rates of the error bounds of the standard method. However, the result does not say that these values of r are the only values that will recover the decay rates of the error bounds of the standard method.

Indeed, for the eigenfunction errors we can choose any $r \leq -1$ and substitute $\Delta = G^r$ into $\delta(M, \widetilde{\mathcal{M}}_{G,\Delta}) \lesssim \Delta^{3/2} + G^{-3/2+\epsilon}$ (from (4.37)) to get eigenfunction errors that are $\mathcal{O}(G^{-3/2+\epsilon})$ for any $\epsilon > 0$, i.e. by choosing *any* $r \leq -1$ we have recovered the eigenfunction error decay rate for the standard method.

For the eigenvalue errors there are also many choices of r that will recover the convergence rate from the standard method. If we choose $r \leq -2$ and substitute $\Delta = G^r$ into $|\tilde{\lambda} - \tilde{\lambda}_j| \lesssim \Delta^{3/2} + G^{-3+2\epsilon}$ (from (4.38)) then we get an eigenvalue error that is $\mathcal{O}(G^{-3+\epsilon})$ for any $\epsilon > 0$, i.e. by choosing *any* $r \leq -2$ we can recover the eigenvalue error convergence rate for the standard method.

Now we realise that these choices of r all correspond to choosing very small Δ, and when we choose very small Δ the errors behave in the same way as the standard method. It is as if we have chosen Δ so small that the method does not recognise that there is any smoothing at all.

This concludes our theoretical error convergence analysis for the smooth problem. However, we mention that all of the above results are also true for the 1D problem with very similar proofs but they are omitted from this thesis.

We now compute some numerical examples to test our theory.

4.3.3 Examples

In this subsection we present numerical examples that support the theoretical results we have developed for solving the smooth problem. We solve Model Problems 1-4 from Section 4.1.7 using the method we have described in this section for $\Delta \neq 0$ and varying

G, and for varying Δ with fixed G. We then implement various strategies to balance the errors by choosing $\Delta = G^r$ for different constants r.

In Figures 4-11 to 4-14 we have plotted the errors of the Galerkin method applied to the smooth problem (Problem 4.28) for fixed Δ and varying G for Model Problems 1-4. For Problems 1 and 2 we have fixed $\Delta = 10^{-4}$ and in Problems 3 and 4 we have fixed $\Delta = 10^{-2}$. The reference solution, which should be the solution to Problem 4.27, is the computed solution to Problem 4.28 with $\Delta = 10^{-4}$ and $G = 2^{18} - 1$ for Problems 1 and 2 and $\Delta = 10^{-2}$ and $G = 2^{10} - 1$ for Problems 3 and 4. Theorem 4.33 implies that we should observe algebraic convergence with respect to G of arbitrary degree for both the eigenvalue and eigenfunction, i.e. superalgebraic convergence. This is indeed what we observe in Figures 4-11 - 4-14 before the error tolerance of the computed reference solutions are reached. However, Theorem 4.33 is an asymptotic result and in some of the plots the faster convergence only occurs for larger G.

In Figures 4-15 - 4-18 we plot the error of Problem 4.27 with respect to the solution of Problem 4.6 for varying Δ. We do not have the exact solutions for these problems so we approximate their solutions by solving Problems 4.17 and 4.28 with large G ($2^{18} - 1$ for the 1D problems and $2^{10} - 1$ for the 2D problems) to get our reference solution and the solution to Problem 4.27 for varying Δ. Theorem 4.31 implies that the eigenvalue and eigenfunction errors should converge with rate $\Delta^{3/2}$. We see that this is indeed the case for the eigenfunctions in all of the model problems. However, for the eigenvalue errors, we observe that our theory is not completely sharp. The eigenvalue errors appear to actually converge with rate Δ^2.

Given this new (numerically observed) rate of convergence for the eigenvalue error of Problem 4.27, we can redo the optimisation for the eigenvalue error in Corollary 4.35 to check whether this changes our conclusion that "no amount of smoothing will give faster convergence than the standard method". We find that based on the numerically observed rate of convergence for the eigenvalue error, the optimal choice for r is $r = -3/2$ (actually, we could choose any $r \leq -3/2$ and get the same rate of convergence). This gives an error bound of the form

$$|\tilde{\lambda} - \tilde{\lambda}_j| \lesssim G^{-3+2\epsilon} \qquad \text{for } \epsilon > 0$$

and $j = 1, \ldots, m$, which is again not faster than the rate of decay of the error bound for the standard method in Section 4.2. Therefore, our conclusion based on numerical observations is the same, "No choice of smoothing will result in a rate of convergence that is faster than the standard method".

Finally, we plot the errors of Problem 4.28 for varying G where we have chosen $\Delta = G^r$ for different values of r. We plot the 1st eigenvalue error from Model Problems 1 and 2 in Figure 4-19 and the 1st eigenvalue error from Model Problem 3 and 4 in Figure 4-20. The 1st eigenfunction errors for Model Problems 1-4 are plotted in

Figures 4-21 and 4-22. The reference solution is Problem 4.17 with $G = 2^{18} - 1$ for the 1D problems and $G = 2^{10} - 1$ for the 2D problems. As well as plotting errors for $\Delta = G^{-1/2}$, $\Delta = G^{-1}$ and $\Delta = G^{-3/2}$ we have also plotted the case when $\Delta = 0$ for comparison. The $\Delta = 0$ case corresponds to the standard method of Section 4.2. In all of the plots we observe that the error convergence rate is never better than the convergence rate of the standard method. We also observe that our optimal choice of smoothing from Corollary 4.35 and the discussion in the previous paragraph ($r = -1$ for eigenfunctions and $r = -3/2$ for eigenvalues) corresponds to the largest choice of Δ (i.e. largest amount of smoothing) that can be chosen without the error converging at a slower rate than the standard method. We interpret this as, "if the amount of smoothing is too big, then the error from smoothing is larger than the error from the plane wave approximation".

To reiterate our conclusion, there is no choice of smoothing that will improve the rate of convergence so that the smoothing method performs better than the standard method. However, we can apply smoothing, up to a point, without having a detrimental effect on the rate of convergence.

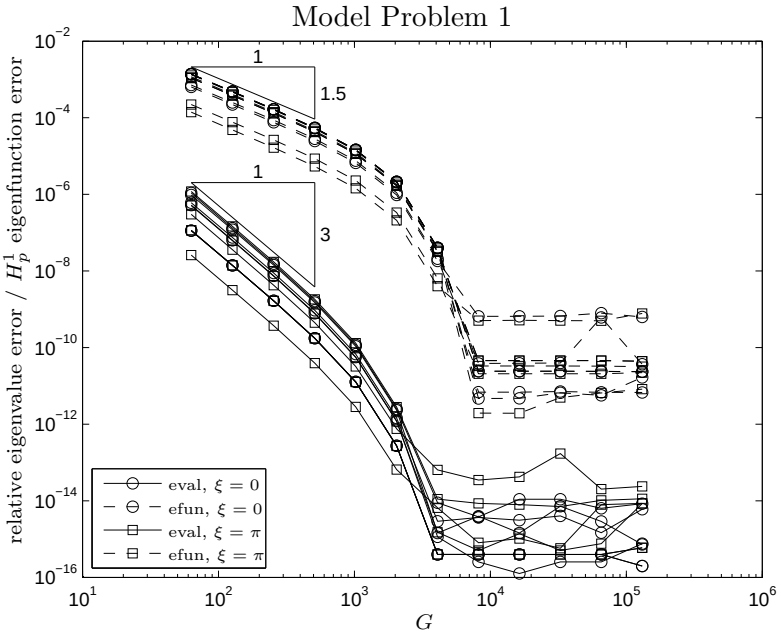

Figure 4-11: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. G for the 1st 5 eigenpairs of Problem 4.28 with $\Delta = 10^{-4}$ fixed.

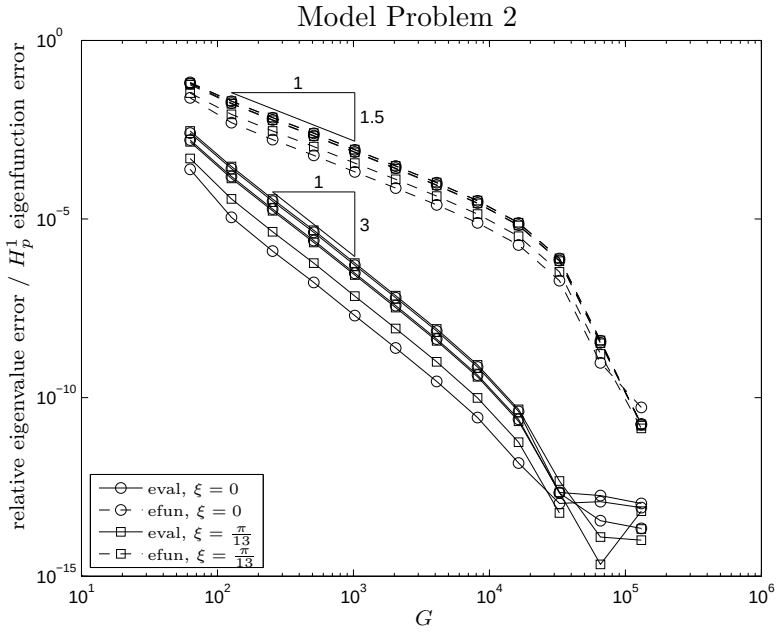

Figure 4-12: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. G for the 37-39th eigenpairs of Problem 4.28 with $\Delta = 10^{-4}$ fixed.

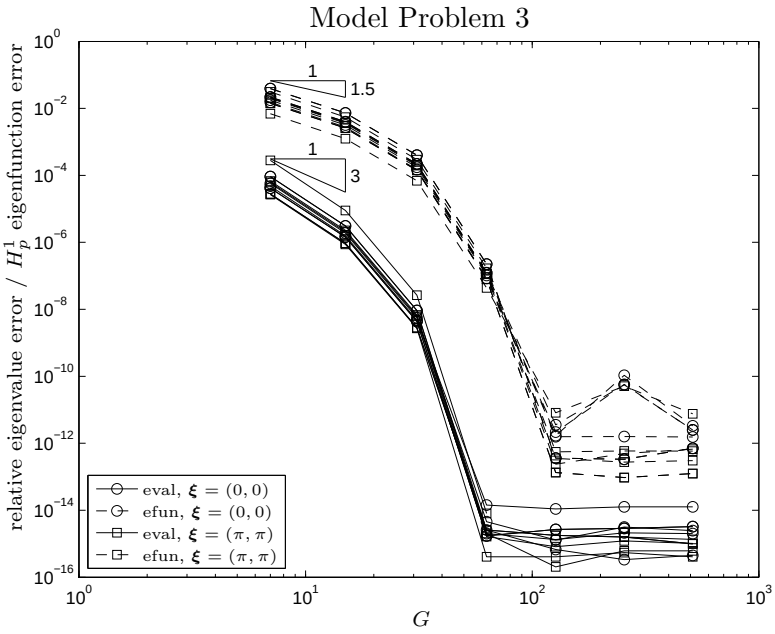

Figure 4-13: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. G for the first 5 eigenpairs of Problem 4.28 with $\Delta = 10^{-2}$ fixed.

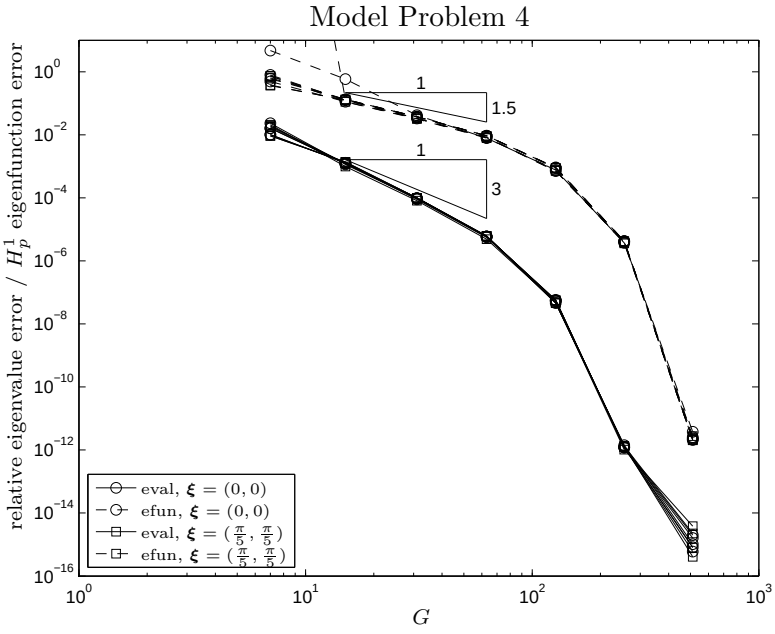

Figure 4-14: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. G for the 23-27th eigenpairs of Problem 4.28 with $\Delta = 10^{-2}$ fixed.

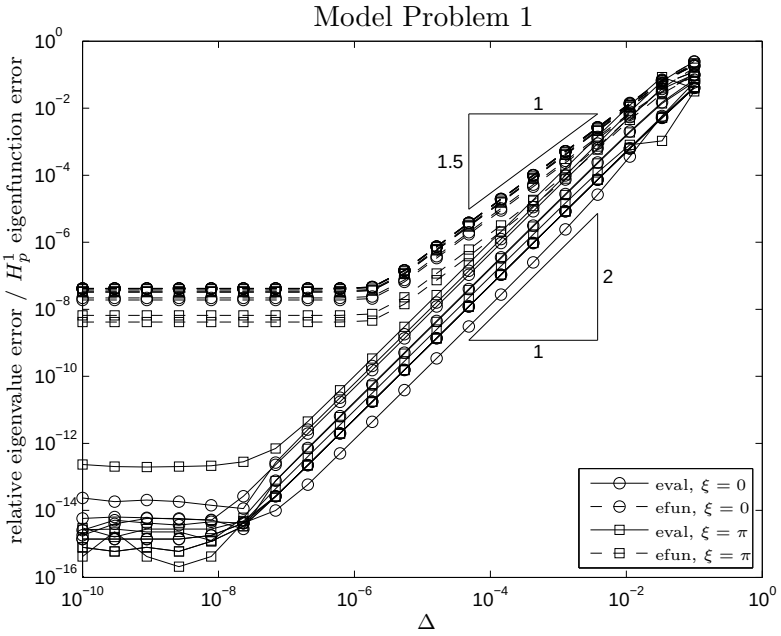

Figure 4-15: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. Δ for the 1st 5 eigenpairs of Problem 4.28 with $G = 2^{16} - 1$ fixed.

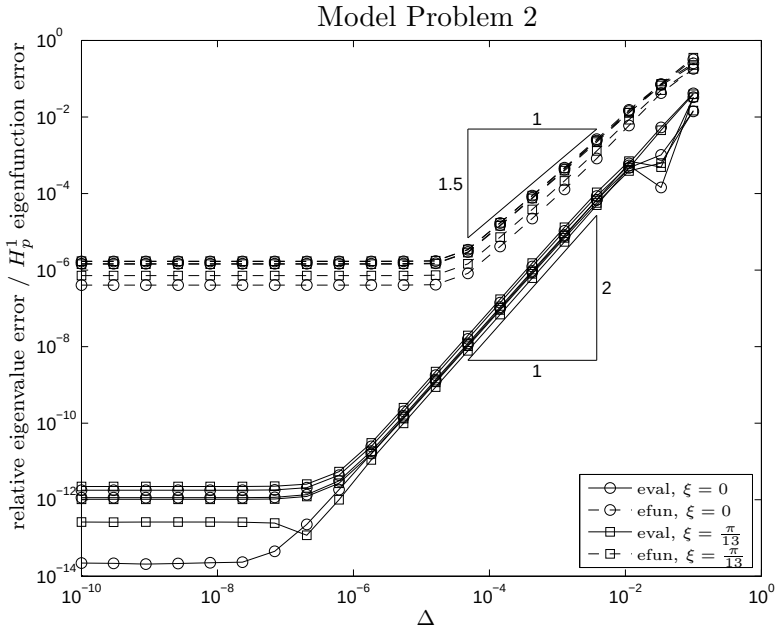

Figure 4-16: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. Δ for the 37-39th eigenpairs of Problem 4.28 with $G = 2^{16} - 1$ fixed.

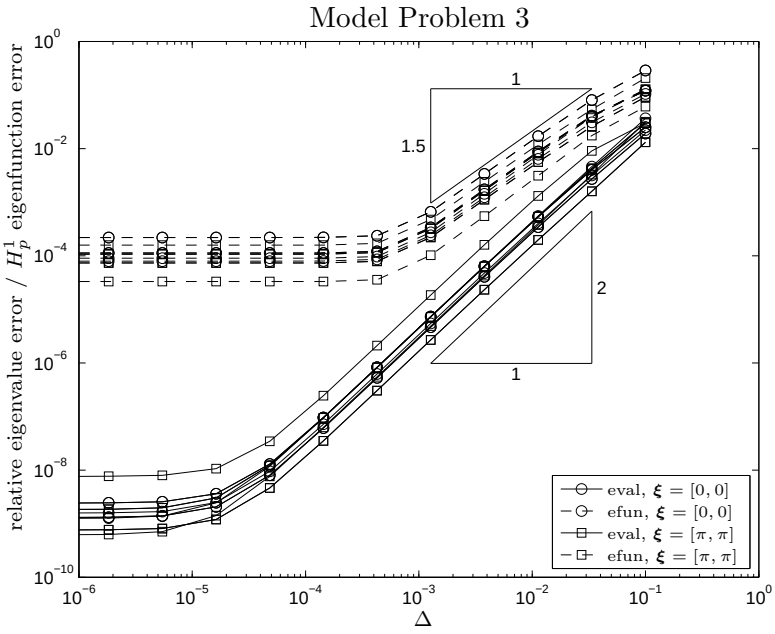

Figure 4-17: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. Δ for the 1st 5 eigenpairs of Problem 4.28 with $G = 2^8 - 1$ fixed.

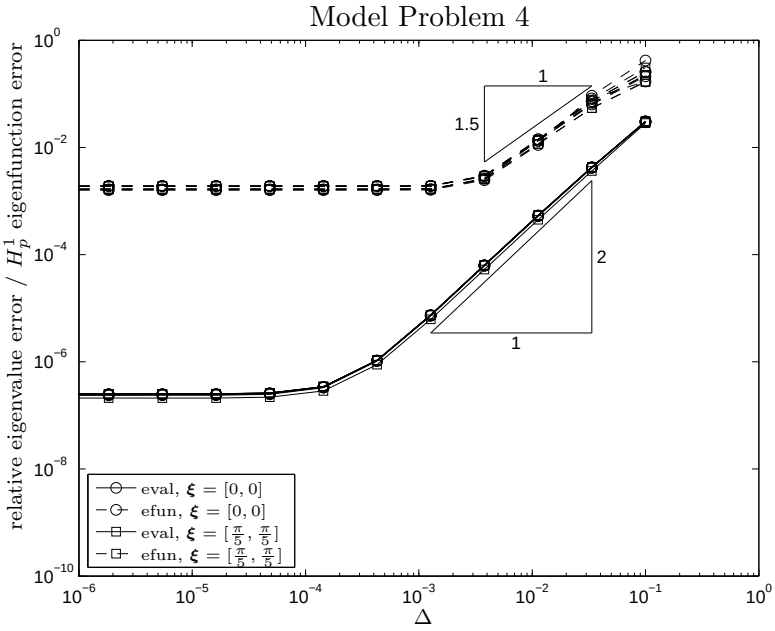

Figure 4-18: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. Δ for the 23-27th eigenpairs of Problem 4.28 with $G = 2^8 - 1$ fixed.

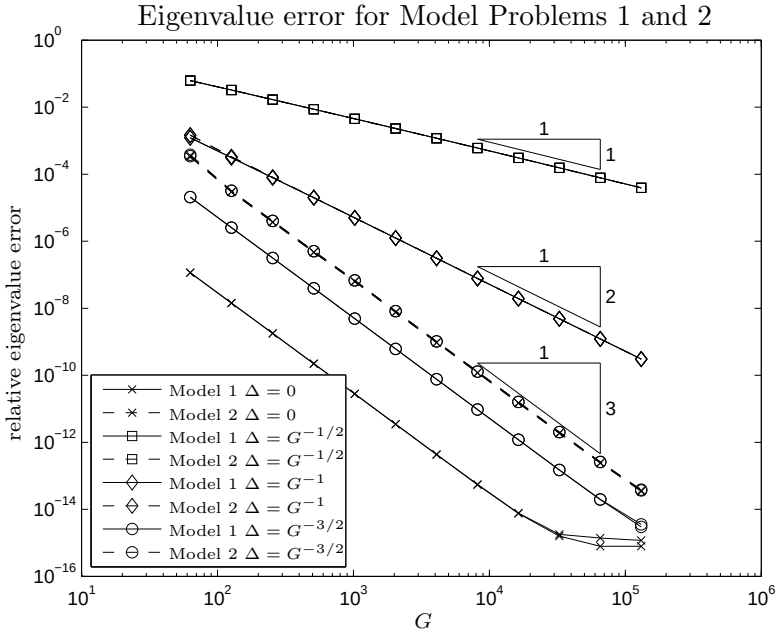

Figure 4-19: Plot of the relative error vs. G for the 1st eigenvalue of Problem 4.28 for $\xi = 0$, and $\xi = \pi$ (for Model Problem 1) or $\xi = \frac{\pi}{13}$ (for Model Problem 2). Note that machine accuracy is reached for the $\Delta = 0$ case for large G.

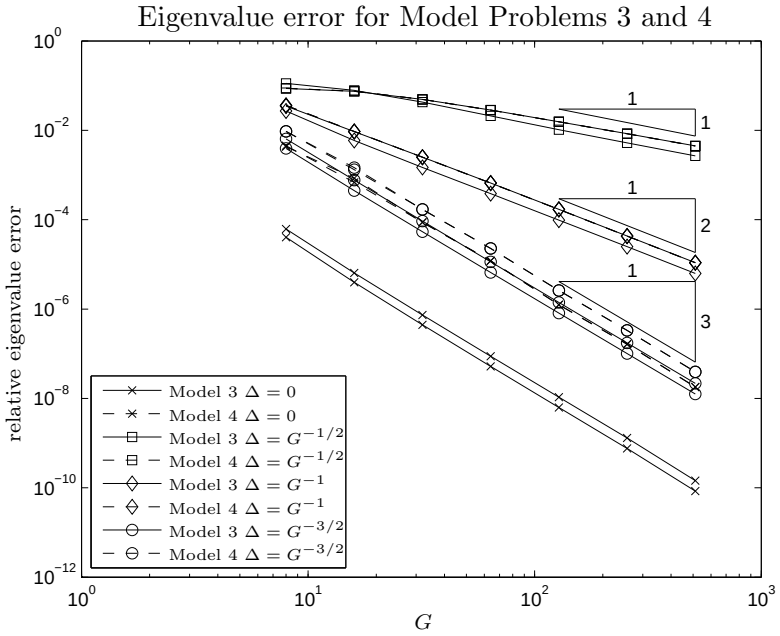

Figure 4-20: Plot of the relative error vs. G for the 1st eigenvalue of Problem 4.28 for $\boldsymbol{\xi} = (0, 0)$, and $\boldsymbol{\xi} = (\pi, \pi)$ (for Model Problem 3) or $\boldsymbol{\xi} = (\frac{\pi}{5}, \frac{\pi}{5})$ (for Model Problem 4).

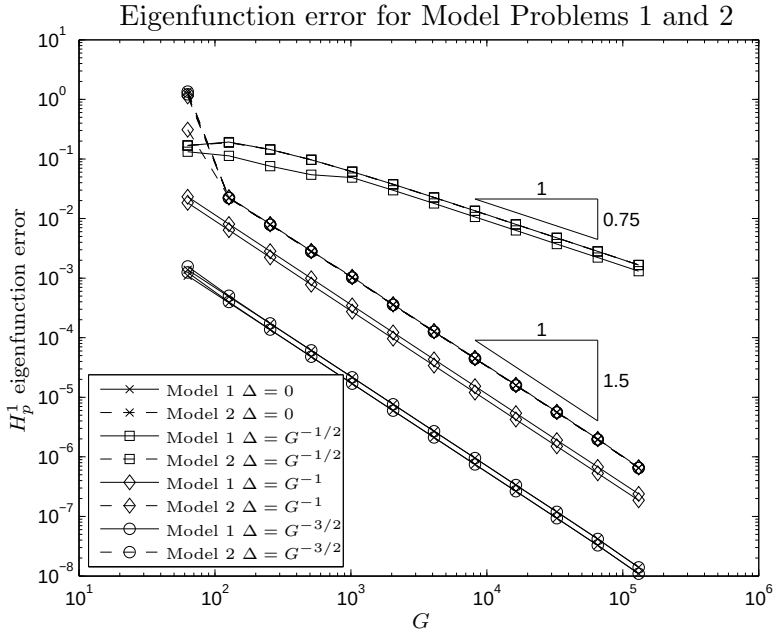

Figure 4-21: Plot of the H_p^1 norm of the error vs. G for the 1st eigenfunction of Problem 4.28 for $\xi = 0$, and $\xi = \pi$ (for Model Problem 1) or $\xi = \frac{\pi}{13}$ (for Model Problem 2).

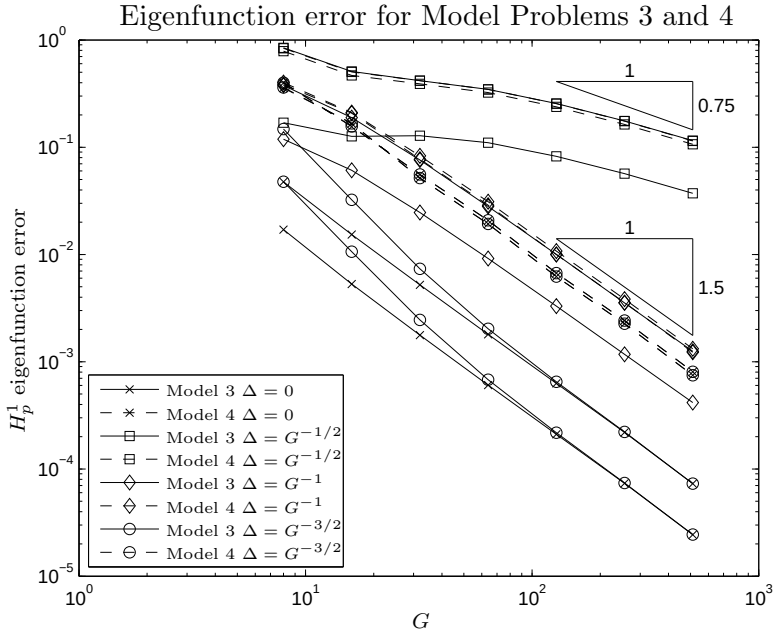

Figure 4-22: Plot of the H_p^1 norm of the error vs. G for the 1st eigenfunction of Problem 4.28 for $\boldsymbol{\xi} = (0, 0)$, and $\boldsymbol{\xi} = (\pi, \pi)$ (for Model Problem 3) or $\xi = (\frac{\pi}{5}, \frac{\pi}{5})$ (for Model Problem 4).

4.4 Sampling

In practice, for more complicated $\gamma(\mathbf{x}) \in PC_p$, we may not have an explicit formula for the Fourier coefficients of $\gamma(\mathbf{x})$. In this case we do not know the entries of the matrix A in (4.14), or equivalently, we do not have the input values for Algorithm 4.19 to compute the action of matrix-vector multiplication with A. So far in this chapter we have assumed that we have an explicit formula for the Fourier coefficients of $\gamma(\mathbf{x})$. Let us now consider the case where we do not have an explicit formula, and we must somehow approximate the Fourier coefficients of $\gamma(\mathbf{x})$.

In this section we make the assumption that $\gamma \in PC_p'$ so that we can apply 3.47 when $d = 2$.

In the first subsection we present a fast and efficient method that utilises the Fast Fourier Transform (FFT) for approximating the Fourier coefficients of $\gamma(\mathbf{x})$. We call this new method the *sampling method*. In the second subsection we analyse the additional error that the sampling method introduces and in the final subsection we present some examples to support our theoretical results.

4.4.1 The method

In this subsection we define the *sampling method* for solving Problem 4.6 when we do not have an explicit formula for the Fourier coefficients of $\gamma(\mathbf{x})$. As we saw in Algorithm 4.19 we do not need all of the Fourier coefficients of $\gamma(\mathbf{x})$. We only require $[\gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$ where N_f is the number that defines the size of the FFT that is used in Algorithm 4.19. The sampling method is to approximate $[\gamma]_{\mathbf{g}}$ with $[\mathrm{Q}_M \gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$ where Q_M is the interpolation projector described in Subsection 3.2.5 and M is a chosen integer that will determine the accuracy of the sampling method.

The reason that we choose this particular projection of $\gamma(\mathbf{x})$ is because it is very easy and efficient to compute $[\mathrm{Q}_M \gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$. Recall that $\mathrm{Q}_M \gamma \in \mathcal{T}_M^{(2)}$ and so, according to our discussion in Subsection 3.2.4, we can represent $\mathrm{Q}_M \gamma$ as a $M \times M$ matrix of either nodal values on a uniform grid or Fourier coefficients. Moreover, using the FFT will allow us to swap between these two different representations a cost of only $\mathcal{O}(M^2 \log M)$ operations. This is the basis of the sampling method.

First, we represent $\mathrm{Q}_M \gamma$ with a matrix of nodal values by sampling $\gamma(\mathbf{x})$ on a uniform grid. We then compute $[\mathrm{Q}_M \gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{M,\square}$ using the FFT. If $M \geq N_f$ (as is usually the case in practice) then we automatically have $[\mathrm{Q}_M \gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$. However, if $M < N_f$ then we recall that $[\mathrm{Q}_M \gamma]_{\mathbf{g}} = 0$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square} \backslash \mathbb{Z}^2_{M,\square}$. We present this process more formally in the following algorithm.

Algorithm 4.36. Choose $M = 2^n$ for some $n \in \mathbb{N}$. Define $\mathbf{g}_0 = (\frac{N_f}{2} + 1, \frac{N_f}{2} + 1)$ and $\mathbf{m}_0 = (\frac{M}{2} + 1, \frac{M}{2} + 1)$. Let $\mathrm{fft}(\cdot)$ denote the 2D Fast Fourier Transform as defined in Subsection 3.2.4. This algorithm computes $[\mathrm{Q}_M \gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f}$ and stores the values

in a matrix $\widehat{Y}$ where $\widehat{Y}_{ij} = [\mathrm{Q}_M \gamma]_{(i,j)-\mathbf{g}_0}$ for $i, j = 1, \ldots, N_f$.
$W_{ij} \leftarrow \gamma\left(\frac{(i,j)-\mathbf{m}_0}{M}\right)$ for $i, j = 1, \ldots, M$
$\widehat{W} \leftarrow \mathrm{fft}(W)$
if $N_f \leq M$ then
    $\widehat{Y}_{ij} \leftarrow \widehat{W}_{(i,j)+\mathbf{x}_0-\mathbf{g}_0}$ for $i, j = 1, \ldots, N_f$.
else
    $\widehat{Y}_{ij} \leftarrow 0$ for $i, j = 1, \ldots, N_f$.
    $\widehat{Y}_{ij} \leftarrow \widehat{W}_{(i,j)+\mathbf{x}_0-\mathbf{g}_0}$ for $i, j = 1, \ldots, M$.
end if

This is the algorithm we use for the 2D problem. There is a similar algorithm for the 1D problem.

Algorithm 4.36 requires one FFT and the total computational cost of the algorithm is $\mathcal{O}(M^2 \log M)$ operations ($\mathcal{O}(M \log M)$ for the 1D problem). When we use Algorithm 4.36 with Algorithm 4.19 to solve (4.14) we only apply Algorithm 4.36 once, while Algorithm 4.19 is applied many times. For this reason we may choose M significantly larger than N_f without incurring a significant increase to the computational cost of solving (4.14).

The additional memory required for the sampling method is an $M \times M$ complex double matrix.

To see that Algorithm 4.36 for approximating $[\gamma]_{\mathbf{g}}$ is efficient, let us compare it with a quadrature method for approximating $[\gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$. For each $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$ an M^2-point quadrature rule method to approximate the integral

$$\gamma_{\mathbf{g}} = \int_\Omega \gamma(\mathbf{x})\, \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{x}}\, dxdy$$

would require $\mathcal{O}(M^2)$ operations. The total cost of computing $[\gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$ would be $\mathcal{O}(M^2 N_f^2)$. Thus, the $\mathcal{O}(M^2 \log M)$ cost of Algorithm 4.36 compares extremely favourably with using M^2-point quadrature to approximate $[\gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$. The main saving comes from computing all of the approximate Fourier coefficients at once rather than repeating the quadrature rule for each approximate Fourier coefficient.

We must now consider the error associated with approximating $[\gamma]_{\mathbf{g}}$ with $[\mathrm{Q}_M \gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$. To bound the errors for the variational eigenvalue problem we must bound $\|\gamma - \mathrm{Q}_M \gamma\|_{H_p^{-1}}$. To do this we cannot directly apply Lemma 3.32 because we are not sure if $\gamma \in H_p^t$ for some $t > 1$ ($t > 1/2$ in 1D). Therefore, we consider a mollified $\gamma(\mathbf{x})$, $\gamma^\delta(\mathbf{x})$. For small $\delta > 0$ we define $\gamma^\delta(\mathbf{x})$ by

$$\gamma^\delta(\mathbf{x}) := J_\delta * \gamma(\mathbf{x}) = \int_{\mathbb{R}^d} J_\delta(\mathbf{y})\gamma(\mathbf{x}-\mathbf{y})d\mathbf{y} = \int_{\mathbb{R}^d} J_\delta(\mathbf{x}-\mathbf{y})\gamma(\mathbf{y})d\mathbf{y} \qquad \forall \mathbf{x} \in \mathbb{R}^d$$

where $J_\delta(\mathbf{x}) = \delta^{-d} J(\delta^{-1}\mathbf{x})$ and $J(\mathbf{x})$ is the standard mollifier that we defined in Sub-

section 3.1.5. In a lemma that follows, Lemma 4.37, we prove some properties about $\gamma^\delta(\mathbf{x})$.

Also note that, Lemma 3.32 can only provide an upper bound for $\|\gamma^\delta - \mathrm{Q}_M \gamma^\delta\|_{H_p^0}$ and not $\|\gamma^\delta - \mathrm{Q}_M \gamma^\delta\|_{H_p^s}$ with $s < 0$ (in particular $s = -1$) and we will need to use the fact that $\|u\|_{H_p^s} \leq \|u\|_{H_p^t}$ for all $u \in H_p^s$ for any $s < t$. This might be where we loose the sharpness for our error bounds.

When we replace $\gamma(\mathbf{x})$ with $\gamma^\delta(\mathbf{x}) \in C_p^\infty$ we will obtain $\mathrm{Q}_M \gamma = \mathrm{Q}_M \gamma^\delta$ if we choose $\delta > 0$ sufficiently small so that $\gamma(\frac{1}{M}\mathbf{k}) = \gamma^\delta(\frac{1}{M}\mathbf{k})$ for all $\mathbf{k} \in \mathbb{Z}^2_{M,\square}$. However, we cannot choose δ arbitrarily small without penalty. The penalty appears in the form of a negative exponent of δ in Parts 3 and 5 of Lemma 4.37. To alleviate this penalty we define yet another approximation to $\gamma(\mathbf{x})$ that will ensure that we can choose $\delta \propto M^{-1}$.

Associated with Q_M are the nodes, $\{\frac{1}{M}\mathbf{k} : \mathbf{k} \in \mathbb{Z}^d_{M,\square}\}$. For $d = 1$ we construct a mesh of uniform intervals with length $\frac{1}{M}$ and for $d = 2$ we construct a mesh of uniform squares with side length $\frac{1}{M}$ such that each node is the centre of an interval (for $d = 1$) or a square (for $d = 2$). We define a perturbed $\gamma(\mathbf{x})$, $\overline{\gamma}(\mathbf{x})$, such that $\overline{\gamma}(\mathbf{x})$ is constant on each of the intervals or squares in the mesh and $\overline{\gamma}(\mathbf{x})$ is equal to $\gamma(\mathbf{x})$ at the nodes, i.e. $\gamma(\mathbf{x}) = \overline{\gamma}(\mathbf{x})$ for all $\mathbf{x} \in \{\frac{1}{M}\mathbf{k} : \mathbf{k} \in \mathbb{Z}^d_{M,\square}\}$. See Figure 4-23 for an example of how we construct $\overline{\gamma}$ from γ for $d = 2$. In Lemma 4.38 we bound the difference between $\gamma(\mathbf{x})$ and $\overline{\gamma}(\mathbf{x})$ in the L_p^2 norm (which is the same as the $L^2(\Omega)$ norm and is equivalent to the H_p^0 norm). Before we bound $\|\gamma - \mathrm{Q}_M \gamma\|_{H_p^0}$ let us prove some properties for the molified $\gamma(\mathbf{x})$.

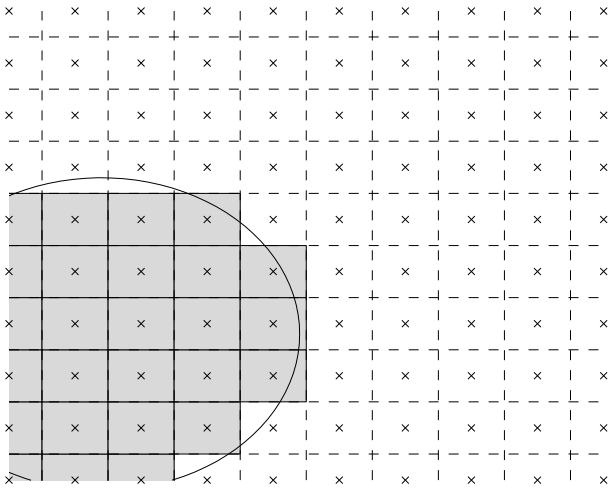

Figure 4-23: Diagram of $\overline{\gamma}$ and γ for $d = 2$. "x" mark the nodes corresponding to Q_M. The dotted lines are the uniform mesh of squares and the grey region is $\overline{\gamma}$. The curved line is an interface of γ.

Lemma 4.37. *Let $d = 1, 2$ and assume $\gamma \in PC'_p$. For $\frac{1}{2} > \delta > 0$ and any $\epsilon > 0$ we get:*

1. $$[\gamma^\delta]_{\mathbf{g}} = [\gamma]_{\mathbf{g}}[J_\delta]_{\mathbf{g}} \qquad \text{for all } \mathbf{g} \in \mathbb{Z}^d.$$

2. $[J_\delta]_0 = 1$, $|[J_\delta]_{\mathbf{g}}| \leq 1$ *for all* $\mathbf{g} \in \mathbb{Z}^d$, *and for any* $k \in \mathbb{N}$,
$$|[J_\delta]_{\mathbf{g}}| \lesssim (\delta|\mathbf{g}|)^{-k} \qquad \text{for all } 0 \neq \mathbf{g} \in \mathbb{Z}^d.$$

3. $$\|\gamma^\delta\|_{H^s_p} \lesssim \begin{cases} 1 & \text{if } s < \frac{1}{2} \\ \delta^{-s+1/2-\epsilon} & \text{if } s \geq \frac{1}{2} \end{cases}.$$

4. $$|[\gamma - \gamma^\delta]_{\mathbf{g}}| \lesssim \begin{cases} 0 & \mathbf{g} = 0 \\ |[\gamma]_{\mathbf{g}}| & \mathbf{g} \in \mathbb{Z}^2 \\ (\delta|\mathbf{g}|)^2|[\gamma]_{\mathbf{g}}| & \mathbf{g} \in \mathbb{Z}^2,\ |g_i| \leq \delta^{-1}. \end{cases}.$$

5. $$\|\gamma - \gamma^\delta\|_{H^s_p} \lesssim \delta^{-s+1/2} \qquad \text{for } -\tfrac{3}{2} < s < \tfrac{1}{2}.$$

Proof. Part 1. From Definition 3.13 we have for $\mathbf{g} \in \mathbb{Z}^d$,

$$\begin{aligned}
[\gamma^\delta]_{\mathbf{g}} &= \int_\Omega \gamma^\delta(\mathbf{x})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \\
&= \int_\Omega \int_{B(0,\delta)} J_\delta(\mathbf{y})\gamma(\mathbf{x}-\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{y} d\mathbf{x} \\
&= \int_\Omega \int_{B(0,\delta)} J_\delta(\mathbf{y}) \left(\sum_{\mathbf{n}\in\mathbb{Z}^d} [\gamma]_{\mathbf{n}}\, \mathrm{e}^{i2\pi\mathbf{n}\cdot(\mathbf{x}-\mathbf{y})} \right) \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{y} d\mathbf{x} \\
&= \sum_{\mathbf{n}\in\mathbb{Z}^d} [\gamma]_{\mathbf{n}} \int_\Omega \mathrm{e}^{i2\pi(\mathbf{n}-\mathbf{g})\cdot\mathbf{x}}\, d\mathbf{x} \int_{B(0,\delta)} J_\delta(\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{n}\cdot\mathbf{y}}\, d\mathbf{y} \\
&= [\gamma]_{\mathbf{g}} \int_{B(0,\delta)} J_\delta(\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{y}}\, d\mathbf{y} \\
&= [\gamma]_{\mathbf{g}} \int_\Omega J_\delta(\mathbf{y})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{y}}\, d\mathbf{y} \\
&= [\gamma]_{\mathbf{g}}[J_\delta]_{\mathbf{g}}.
\end{aligned}$$

Part 2. $[J_\delta]_0 = 1$ follows from the definition of J_δ. For all $\mathbf{g} \in \mathbb{Z}^d$,

$$|[J_\delta]_{\mathbf{g}}| = \left| \int_\Omega J_\delta(\mathbf{x})\, \mathrm{e}^{-i2\pi\mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \right| \leq \int_\Omega J_\delta(\mathbf{x}) d\mathbf{x} = 1.$$

For $0 \neq \mathbf{g} \in \mathbb{Z}^d$, $g_i \neq 0$ and integrating by parts gives us

$$\begin{aligned}
[J_\delta]_{\mathbf{g}} &= \int_\Omega J_\delta(\mathbf{x})\, \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \\
&= \left(\tfrac{-1}{i2\pi g_i}\right)^k \int_\Omega D^k_{x_i} J_\delta(\mathbf{x})\, \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \\
&= \left(\tfrac{-1}{i2\pi g_i}\right)^k \int_\Omega \delta^{-d-k} \left(D^k_{y_i} J(\mathbf{y})\right)\Big|_{\mathbf{y}=\frac{\mathbf{x}}{\delta}} \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x} \\
&= \left(\tfrac{-1}{i2\pi g_i}\right)^k \int_{B(0,\delta)} \delta^{-d-k} \left(D^k_{y_i} J(\mathbf{y})\right)\Big|_{\mathbf{y}=\frac{\mathbf{x}}{\delta}} \mathrm{e}^{-i2\pi \mathbf{g}\cdot\mathbf{x}}\, d\mathbf{x}.
\end{aligned}$$

This implies that

$$|[J_\delta]_{\mathbf{g}}| \leq \left(\tfrac{1}{2\pi g_i \delta}\right)^k \max_{|\alpha|=k} \|D^\alpha J\|_\infty.$$

Now, the result follows from

$$|\mathbf{g}|^k |[J_\delta]_{\mathbf{g}}| \leq d^{k/2} \sum_{i=1}^d |g_i|^k |[J_\delta]_{\mathbf{g}}| \leq \tfrac{d^{k+1/2}}{(2\pi\delta)^k} \max_{|\alpha|=k} \|D^\alpha J\|_\infty.$$

Part 3. We only prove Part 3 for $d = 2$. The proof for $d = 1$ is similar. First consider the case when $s < 1/2$. Using Parts 1 and 2 we get $|[\gamma^\delta]_{\mathbf{g}}| \leq |[\gamma]_{\mathbf{g}}|$ for all $\mathbf{g} \in \mathbb{Z}^d$. Therefore, $\|\gamma^\delta\|_{H^s_p} \leq \|\gamma\|_{H^s_p} \lesssim 1$ by Theorem 3.40.

For $s \geq 1/2$, let $k \in \mathbb{N} \cup \{0\}$, and get

$$\begin{aligned}
\|\gamma^\delta\|^2_{H^s_p} &= \sum_{\mathbf{g}\in\mathbb{Z}^2} |\mathbf{g}|^{2s}_\star |[\gamma^\delta]_{\mathbf{g}}|^2 \\
&\lesssim |[\gamma]_0| + \delta^{-2k} \sum_{0\neq\mathbf{g}\in\mathbb{Z}^2} |\mathbf{g}|^{2s-2k} |[\gamma]_{\mathbf{g}}|^2 && \text{by Parts 1 and 2} \\
&= |[\gamma]_0| + \delta^{-2k} \sum_{n=1}^\infty \sum_{|g_1|+|g_2|=n} |\mathbf{g}|^{2s-2k} |[\gamma]_{\mathbf{g}}|^2 \\
&\lesssim |[\gamma]_0| + \delta^{-2k} \sum_{n=1}^\infty n^{2s-2k} C_n^2 && C_n \text{ from Theorem 3.47} \\
&\lesssim \delta^{-2k} \sum_{n=1}^\infty n^{2s-2k-2} && \text{by Theorem 3.47} \\
&\lesssim \delta^{-2k} \quad \text{provided } s < k + \tfrac{1}{2}.
\end{aligned}$$

Therefore, $\|\gamma^\delta\|_{H^s_p} \lesssim \delta^{-k}$ provided $s < k + 1/2$. The result follows by using Lemma 3.26 with (s from Lemma 3.26) $s = k + 1/2 - \epsilon$ and $t = k + 3/2 - \epsilon$.

Part 4. We do this proof for $d = 2$. The argument for $d = 1$ is similar and easier, and so we omit it. Part 2 gives us $[J_\delta]_0 = 1$. This together with Part 1 imply that $|[\gamma - \gamma^\delta]_0| = 0$. Also, it follows from Parts 1 and 2 that $|[\gamma - \gamma^\delta]_{\mathbf{g}}| \leq 2|[\gamma]_{\mathbf{g}}|$ for all $\mathbf{g} \in \mathbb{Z}^2$.

For $0 \neq \mathbf{g} \in \mathbb{Z}^2$ we can also get

$$\begin{aligned}
|[\gamma - \gamma^\delta]_{\mathbf{g}}| = |[\gamma]_{\mathbf{g}} - [\gamma^\delta]_{\mathbf{g}}| &= |[\gamma]_{\mathbf{g}}| \left| \int_{B(0,1)} J(\mathbf{x}) \left(1 - \mathrm{e}^{-i2\pi\delta \mathbf{g}\cdot\mathbf{x}}\right) d\mathbf{x} \right| \quad \text{by Part 1} \\
&\leq |[\gamma]_{\mathbf{g}}| \left| \int_{-1}^{1} \int_{-1}^{1} J(\mathbf{x}) \left(1 - \cos(2\pi\delta g_1 x_1)\cos(2\pi\delta g_2 x_2)\right) dx_1 dx_2 \right| \\
&\leq |[\gamma]_{\mathbf{g}}| \|J\|_\infty \int_{-1}^{1} \int_{-1}^{1} \left(1 - \cos(2\pi\delta g_1 x_1)\cos(2\pi\delta g_2 x_2)\right) dx_1 dx_2 \\
&= |[\gamma]_{\mathbf{g}}| 4 \|J\|_\infty \left(1 - \tfrac{\sin(2\pi\delta g_1)}{2\pi\delta g_1} \tfrac{\sin(2\pi\delta g_2)}{2\pi\delta g_2}\right)
\end{aligned}$$

Note that the third line follows from the second line above because the imaginary integral is 0, since sine is odd and J is even. If $A^2 \leq 42$, then $\frac{A^4}{5!} - \frac{A^6}{7!} \geq 0$ and

$$\tfrac{\sin A}{A} = 1 - \tfrac{A^2}{3!} + \left(\tfrac{A^4}{5!} - \tfrac{A^6}{7!}\right) + \cdots \geq 1 - \tfrac{A^2}{6}.$$

Using this inequality it follows that

$$\begin{aligned}
|[\gamma - \gamma^\delta]_{\mathbf{g}}| &\leq |[\gamma]_{\mathbf{g}}| 4 \|J\|_\infty \left(1 - \left(1 - \tfrac{(2\pi\delta g_1)^2}{6}\right)\left(1 - \tfrac{(2\pi\delta g_2)^2}{6}\right)\right) \\
&\leq |[\gamma]_{\mathbf{g}}| 4 \|J\|_\infty \tfrac{(2\pi\delta g_1)^2 + (2\pi\delta g_1)^2}{6} \\
&= \tfrac{16\|J\|_\infty \pi^2}{3} \delta^2 |\mathbf{g}|^2 |[\gamma]_{\mathbf{g}}| \qquad \text{if } |g_i| \leq \delta^{-1}.
\end{aligned}$$

Part 5. Finally, we prove Part 5 for the $d = 2$ case. Let $-\frac{3}{2} < s < \frac{1}{2}$. Using Part 4, Lemma 3.47 and Lemma 3.9 we get

$$\begin{aligned}
\|\gamma - \gamma^\delta\|_{H_p^s}^2 &= \sum_{0 \neq \mathbf{g} \in \mathbb{Z}_2} |\mathbf{g}|^{2s} |[\gamma - \gamma^\delta]_{\mathbf{g}}|^2 \\
&\lesssim \sum_{|g_1|+|g_2| \leq \lfloor \delta^{-1} \rfloor} \delta^4 |\mathbf{g}|^{2s+4} |[\gamma]_{\mathbf{g}}|^2 + \sum_{|g_1|+|g_2| \geq \lceil \delta^{-1} \rceil} |\mathbf{g}|^{2s} |[\gamma]_{\mathbf{g}}|^2 \\
&\leq \delta^4 \sum_{n=1}^{\lfloor \delta^{-1} \rfloor} n^{2s+4} C_n^2 + \sum_{n=\lceil \delta^{-1} \rceil}^{\infty} n^{2s} C_n^2 \qquad C_n \text{ from Theorem 3.47} \\
&\lesssim \delta^4 \sum_{n=1}^{\lfloor \delta^{-1} \rfloor} n^{2s+2} + \sum_{n=\lceil \delta^{-1} \rceil}^{\infty} n^{2s-2} \qquad \text{by Theorem 3.47} \\
&\leq \delta^4 \left(1 + \int_1^{\delta^{-1}} x^{2s+2} dx + \delta^{-2-2s}\right) + \left(\delta^{2-2s} + \int_{\delta^{-1}}^{\infty} x^{2s-2} dx\right) \\
&= \delta^4 + \frac{1}{2s+3}(\delta^{1-2s} - \delta^4) + \delta^{2-2s} + \delta^{2-2s} + \frac{1}{1-2s}\delta^{1-2s} \\
&\lesssim \delta^{1-2s}.
\end{aligned}$$

The result follows by taking the square root of this expression. □

In Part 5 of the preceding Lemma we have restricted ourselves to the case when $-\frac{3}{2} < s < \frac{1}{2}$. Note, however, that although it is strictly necessary to have $s < \frac{1}{2}$, we may in fact choose $s < -\frac{3}{2}$. We do not include this case because $\|\gamma - \gamma^\delta\|_{H_p^s}$ does not depend on s for $s < \frac{3}{2}$ and the result would be $\|\gamma - \gamma^\delta\|_{H_p^s} \lesssim \delta^2$.

We now prove a lemma that bounds the difference between γ and $\overline{\gamma}$ in the L_p^2 norm. This will be sufficient for our purposes.

Lemma 4.38. *Let $d = 1, 2$. For $\gamma \in PC_p'$, $M \in \mathbb{N}$ and with $\overline{\gamma}(\mathbf{x})$ defined in the discussion before Lemma 4.37 we get*

$$\|\gamma - \overline{\gamma}\|_{L_p^2} \lesssim M^{-1/2}$$

Proof. We first consider the $d = 1$ case. Let J_Ω denote the number of intervals Ω_j in $\gamma(\mathbf{x})$. Therefore, there are $2J_\Omega$ jumps in $\gamma(\mathbf{x})$. At each jump there is a potential difference between $\gamma(\mathbf{x})$ and $\overline{\gamma}(\mathbf{x})$. The size of the difference is bounded by γ_{max}, and for each jump the area in Ω where $\gamma(\mathbf{x})$ and $\overline{\gamma}(\mathbf{x})$ are different is limited to an interval of size M^{-1}. Therefore, we get

$$\|\gamma - \overline{\gamma}\|_{L_p^2} = \left(\int_\Omega |\gamma - \overline{\gamma}|^2 dx\right)^{1/2} \leq \sqrt{2J}\gamma_{max} M^{-1/2}.$$

For $d = 2$ there are $\mathcal{O}(M)$ possible squares where γ is different from $\overline{\gamma}$ since there are finitely many Ω_j and each Ω_j is convex. Again, the size of the difference between γ and $\overline{\gamma}$ is bounded by γ_{max} and each square has area M^{-2}. Therefore, we get

$$\|\gamma - \overline{\gamma}\|_{L_p^2} = \left(\int_\Omega |\gamma - \overline{\gamma}|^2 d\mathbf{x}\right)^{1/2} \lesssim \left(M\gamma_{max} M^{-2}\right)^{1/2} \lesssim M^{-1/2}.$$

□

Now we can (finally) bound the difference between γ and $\mathrm{Q}_M \gamma$.

Lemma 4.39. *Let $d = 1, 2$, $\gamma \in PC_p'$ and $\epsilon > 0$. Then*

$$\|\gamma - \mathrm{Q}_M \gamma\|_{L_p^2} \lesssim M^{-1/2+\epsilon}. \tag{4.42}$$

Proof. For this proof we would like to apply Lemma 3.32, but we are not sure that $\gamma \in H_p^t$ for $t > 1$ if $d = 2$, or $t > 1/2$ if $d = 1$. Instead we could try applying Lemma 3.32 to γ^δ for small $\delta > 0$. But choosing δ too small will not work because the bound will depend on δ^s for some $s < 0$. To avoid having to take very small δ we will apply Lemma 3.32 to $\overline{\gamma}^\delta$ with $\delta = \frac{1}{2M}$. With this choice of δ we have $\gamma(\mathbf{x}) = \overline{\gamma}(\mathbf{x}) = \overline{\gamma}^\delta(\mathbf{x})$ for all $\mathbf{x} \in \{\frac{1}{M}\mathbf{k} : \mathbf{k} \in \mathbb{Z}_{M,\square}^d\}$ as well as being able to apply Lemma 3.32. Since we have equality at the nodes, $\mathrm{Q}_M \gamma = \mathrm{Q}_M \overline{\gamma} = \mathrm{Q}_M \overline{\gamma}^\delta$.

Using the triangle inequality we split (4.42) into the following,

$$\|\gamma - \mathrm{Q}_M\,\gamma\|_{H_p^s} \leq \underbrace{\|\gamma - \overline{\gamma}\|_{H_p^s}}_{I_1} + \underbrace{\|\overline{\gamma} - \overline{\gamma}^\delta\|_{H_p^s}}_{I_2} + \underbrace{\|\overline{\gamma}^\delta - \mathrm{Q}_M\,\overline{\gamma}^\delta\|_{H_p^s}}_{I_3}$$

We use Lemma 4.38 to bound I_1 and we obtain

$$I_1 = \|\gamma - \overline{\gamma}\|_{H_p^s} \lesssim M^{-1/2} \qquad \text{for } s \leq 0. \tag{4.43}$$

To bound I_2 we use Part 5 of Lemma 4.37. Note that $\overline{\gamma} \in H^{1/2-\epsilon}$ for any $\epsilon > 0$ and we get

$$I_2 = \|\overline{\gamma} - \overline{\gamma}^\delta\|_{H_p^s} \lesssim \delta^{-s+1/2} \lesssim M^{-1/2+s} \qquad \text{for } -\tfrac{3}{2} < s < \tfrac{1}{2} \tag{4.44}$$

since $\delta = \frac{1}{2M}$.

To bound I_3 we use Lemma 3.32 and Part 3 of Lemma 4.37 to get (with $t > 1$ for $d = 2$ and $t > 1/2$ for $d = 1$),

$$\begin{aligned} I_3 &= \|\overline{\gamma}^\delta - \mathrm{Q}_M\,\overline{\gamma}^\delta\|_{H_p^s} \\ &\lesssim M^{s-t}\|\overline{\gamma}^\delta\|_{H_p^t} \\ &\lesssim M^{s-t}\delta^{-t+1/2-\epsilon} \\ &\lesssim M^{-1/2+s+\epsilon} \qquad \text{for } s \geq 0 \text{ since } \delta = \tfrac{1}{2M}. \end{aligned} \tag{4.45}$$

Finally, putting together (4.43) - (4.45) with $s = 0$ gives us the result. □

4.4.2 Error Analysis

In this subsection we derive theoretical error bounds for the additional error that we introduce when we use the sampling method with the spectral Galerkin method to approximate the solution to Problem 4.6. As in the previous sections we will define the discrete problem that our method is actually solving and we define the corresponding solution operator for this discrete problem. We then prove some properties of the new solution operator, including bounding the difference between the new solution operator and the solution operator that corresponds to Problem 4.6 in terms of G and M (our sampling parameter). We can then apply Theorem 3.68 to get eigenfunction and eigenvalue error bounds.

The bound for the difference between the solution operators is proved using Strang's 1st Lemma (Theorem 3.75). Unlike the analysis of the smoothing method in Section 4.3 we will not define an intermediate problem and then add two error contributions together. Instead we will bound the error all in one go. We do this because we do not expect (and therefore do not attempt to prove) that the sampling method will improve the the performance of the planewave expansion method.

Throughout this section we assume that $\gamma \in PC'_p$.

By approximating the Fourier coefficients of $\gamma(\mathbf{x})$ with the sampling method, the discrete problem we actually solve is,

Problem 4.40. Find $\lambda_G \in \mathbb{R}$ and $0 \neq u_G \in \mathcal{S}_G$ such that

$$a_{\mathrm{Q}}(u_G, v_G) = \lambda_G b(u_G, v_G) \qquad \forall v_G \in \mathcal{S}_G. \tag{4.46}$$

where

$$a_{\mathrm{Q}}(u, v) = \int_\Omega (\nabla + i\xi)\, u \cdot \overline{(\nabla + i\xi)\, v} + (\mathrm{K} - \mathrm{Q}_M\, \gamma)\, u\overline{v} dx$$

Using very similar proofs to Lemma 4.7 we have that $a_{\mathrm{Q}}(\cdot,\cdot)$ is a bounded, coercive and Hermitian bilinear form and therefore $a_{\mathrm{Q}}(\cdot,\cdot)$ also defines an inner product on $H^1_p(\Omega)$ with an induced norm $\|\cdot\|_{a_{\mathrm{Q}}} := a_{\mathrm{Q}}(\cdot,\cdot)^{1/2}$. We may now define a solution operator corresponding to Problem 4.40 as well as proving some properties for our new solution operator.

Lemma 4.41. *Let $\gamma \in PC'_p$. Problem 4.40 has a corresponding solution operator, $\mathrm{T}_Q(G, M)$, that is defined according to Definition 3.70. $\mathrm{T}_Q : H^1_p \to H^1_p$ is bounded and compact, and self-adjoint with respect to $a_{\mathrm{Q}}(\cdot,\cdot)$, (but not self-adjoint with respect to $a(\cdot,\cdot)$ in general). For sufficiently large G and M, and small $\epsilon > 0$ we get:*

1. $$\|\mathrm{T} - \mathrm{T}_Q\|_{H^1_p} \lesssim G^{-3/2+\epsilon} + M^{-1/2+\epsilon}.$$

2. *The adjoint T^*_Q of T_Q with respect to $a(\cdot,\cdot)$ satisfies*
$$\left\|\mathrm{T} - \mathrm{T}^*_Q\right\|_{H^1(\Omega)} \lesssim G^{-3/2+\epsilon} + M^{-1/2+\epsilon}.$$

3. *For u, v eigenfunctions of Problem 4.6 we get*
$$|a((\mathrm{T} - \mathrm{T}_Q)u, v)| \lesssim (G^{-3+2\epsilon} + M^{-1/2+\epsilon})\|u\|_{H^1_p}\|v\|_{H^1_p}.$$

Proof. Using similar proofs to those given in Lemma 4.23 we can show that $\mathrm{T}_Q : H^1_p \to H^1_p$ is bounded and compact, and self-adjoint with respect to $a_{\mathrm{Q}}(\cdot,\cdot)$.

The proof of Part 1 relies on Strang's 1st Lemma (Theorem 3.75). For $f \in H^1_p$ and

$\epsilon > 0$ we get

$$\begin{aligned}
\|\,\mathrm{T}f - \mathrm{T}_Q\, f\|_{H_p^1} &\lesssim \inf_{v_G\in\mathcal{S}_G}\left\{\|\,\mathrm{T}f - v_G\|_{H_p^1} + \sup_{w_G\in\mathcal{S}_G}\frac{|a(v_G,w_G)-a_{\mathrm{Q}}(v_G,w_G)|}{\|w_G\|_{H_p^1}}\right\}\\
&\le \|\,\mathrm{T}f-\nu\|_{H_p^1} + \sup_{w_G\in\mathcal{S}_G}\frac{|a(\nu,w_G)-a_{\mathrm{Q}}(\nu,w_G)|}{\|w_G\|_{H_p^1}} \quad \text{where } \nu = \mathrm{P}_{\mathcal{S}_G}\mathrm{T}f\\
&\le \frac{\|\,\mathrm{T}f\|_{H_p^{5/2-\epsilon}}}{G^{3/2-\epsilon}} + \sup_{w_G\in\mathcal{S}_G}\frac{\int_\Omega|(\mathrm{Q}_M\,\gamma-\gamma)\nu\overline{w_G}|dx}{\|w_G\|_{H_p^1}} \quad \text{by Lemma 3.30}\\
&\le \frac{\|\,\mathrm{T}f\|_{H_p^{5/2-\epsilon}}}{G^{3/2-\epsilon}} + \sup_{w_G\in\mathcal{S}_G}\frac{\|\nu\|_\infty\int_\Omega|(\mathrm{Q}_M\,\gamma-\gamma)\overline{w_G}|dx}{\|w_G\|_{H_p^1}}\\
&\lesssim \frac{\|\,\mathrm{T}f\|_{H_p^{5/2-\epsilon}}}{G^{3/2-\epsilon}} + \|\nu\|_{H_p^2}\|\,\mathrm{Q}_M\,\gamma-\gamma\|_{H_p^{-1}} \quad \text{by Theorem 3.27}\\
&\le \frac{\|\,\mathrm{T}f\|_{H_p^{5/2-\epsilon}}}{G^{3/2-\epsilon}} + \|\,\mathrm{T}f\|_{H_p^2}\|\,\mathrm{Q}_M\,\gamma-\gamma\|_{H_p^0} \qquad (4.47)\\
&\lesssim \frac{\|\,\mathrm{T}f\|_{H_p^{5/2-\epsilon}}}{G^{3/2-\epsilon}} + \frac{\|\,\mathrm{T}\,f\|_{H_p^2}}{M^{1/2-\epsilon}} \quad \text{by Lemma 4.39}\\
&\lesssim \left(\frac{1}{G^{3/2-\epsilon}}+\frac{1}{M^{1/2-\epsilon}}\right)\|f\|_{H_p^1} \quad \text{by Theorem 4.11.}
\end{aligned}$$

That concludes Part 1.

The proof of Part 2 is identical to the proof of Part 2 in Lemma 4.30. To get the result of Part 3, let $u, v \in H_p^1$ and let $\nu = \mathrm{P}_G^{(\mathcal{S})}\, v$. Then

$$\begin{aligned}
|a((\mathrm{T}-\mathrm{T}_Q)u,v)| \le \underbrace{|a((\mathrm{T}-\mathrm{T}_Q)u,v-\nu)|}_{I_1} &+ |a(\mathrm{T}\,u,\nu)-a_{\mathrm{Q}}(\mathrm{T}_Q\,u,\nu)|\\
&+ \underbrace{|a_{\mathrm{Q}}(\mathrm{T}_Q\,u,\nu)-a(\mathrm{T}_Q\,u,\nu)|}_{I_2}. \qquad (4.48)
\end{aligned}$$

By the definition of T and T_Q we get that $a(\mathrm{T}\,u,\nu)-a_{\mathrm{Q}}(\mathrm{T}_Q\,u,\nu)=0$. Now treat I_1 and I_2 separately.

For I_1 we use that $a(\cdot,\cdot)$ is bounded and Part 1 of this Lemma to get

$$\begin{aligned}
I_1 &= |a((\mathrm{T}-\mathrm{T}_Q)u,v-\nu)|\\
&\lesssim \|(\mathrm{T}-\mathrm{T}_Q)u\|_{H_p^1}\|v-\mathrm{P}_{\mathcal{S}_G}v\|_{H_p^1} && a(\cdot,\cdot)\text{ bounded}\\
&\lesssim \left(\frac{1}{G^{3/2-\epsilon}}+\frac{1}{M^{1/2-\epsilon}}\right)\frac{1}{G^{3/2-\epsilon}}\|u\|_{H_p^1}\|v\|_{H_p^{5/2-\epsilon}} && \text{Part 1 \& Lemma 3.30}\\
&\lesssim \left(G^{-3+2\epsilon}+M^{-1/2+\epsilon}\right)\|u\|_{H_p^1}\|v\|_{H_p^1} && \text{Corollary 4.12.}
\end{aligned}$$

For I_2 we do the following,

$$\begin{aligned} I_2 = |a_{\mathrm{Q}}(\mathrm{T}_Q\, u, \nu) - a(\mathrm{T}_Q\, u, \nu)| &= \left| \int_\Omega (\gamma - \mathrm{Q}_M\, \gamma)(\mathrm{T}_Q\, u)\nu \, d\mathbf{x} \right| \\ &\leq \|\nu\|_\infty \int_\Omega |(\gamma - \mathrm{Q}_M\, \gamma)\, \mathrm{T}_Q\, u| \, d\mathbf{x} \\ &\lesssim \|\, \mathrm{P}_{\mathcal{S}_G} v\|_{H_p^2} \|\gamma - \mathrm{Q}_M\, \gamma\|_{H_p^{-1}} \|\, \mathrm{T}_Q\, u\|_{H_p^1} \qquad \text{Theorem 3.27} \\ &\lesssim \|v\|_{H_p^2} \|\gamma - \mathrm{Q}_M\, \gamma\|_{H_p^0} \|u\|_{H_p^1} \qquad \mathrm{T}_Q \text{ bounded} \\ &\lesssim M^{-1/2+\epsilon} \|u\|_{H_p^1} \|v\|_{H_p^1} \qquad \text{Cor.4.12 \& Lem.4.39.} \end{aligned} \tag{4.49}$$

Now we put I_1 and I_2 back into (4.48) to get the result for Part 3. □

In the preceding proof at (4.47) and (4.49), we may have 'thrown away' the sharpness of our bounds when we bounded $\|\gamma - \mathrm{Q}_M\, \gamma\|_{H_p^{-1}}$ with $\|\gamma - \mathrm{Q}_M\, \gamma\|_{H_p^0}$. We did this because we were unable to bound $\|\gamma - \mathrm{Q}_M\, \gamma\|_{H_p^{-1}}$ with a better dependence on M in Lemma 4.39. In the numerical examples later in this section we show that our error bounds are not sharp, and this may be where we are losing the sharpness of our eigenfunction bound.

We now apply Theorem 3.68 to get bounds on the eigenvalue and eigenfunction errors of solving Problem 4.6 with the sampling method. The proof of the following result is analogous to the proof of Theorem 4.24 and it requires Lemma 4.41.

Theorem 4.42. *Let λ be an eigenvalue of Problem 4.6 (with $\gamma \in PC_p'$) with multiplicity m and corresponding eigenspace $\mathcal{M}$. Then, for sufficiently large G and large M there exist m eigenvalues $\lambda_1(G,M), \ldots, \lambda_m(G,M)$ of Problem 4.40 with corresponding eigenspaces $M_1(\lambda_1), \ldots, M_m(\lambda_m)$ and a space*

$$\mathcal{M}_{G,M} := \bigoplus_{j=1}^{m} M_j(\lambda_j)$$

such that for $\epsilon > 0$,

$$\delta(\mathcal{M}, \mathcal{M}_{G,M}) \lesssim G^{-3/2+\epsilon} + M^{-1/2+\epsilon}$$

and

$$|\lambda - \lambda_j| \lesssim G^{-3+2\epsilon} + M^{-1/2+\epsilon} \qquad \text{for } j = 1, \ldots, m.$$

We could now proceed to balance/optimise the errors by devising a method where we choose $M = CG^r$ for a constant C and r. However, in the numerical examples of the next subsection we discover that our error bounds are not sharp with respect to M. Therefore, we will delay our discussion for choosing r until after we observe the actual dependence of the errors on M.

As we have already discussed, the computational cost for using our sampling method is $\mathcal{O}(M^d \log M)$, but the additional cost is only in the "setup", i.e. we only need to

compute one FFT on an $M^d \times M^d$ matrix to compute the approximate the Fourier coefficients of $\gamma(\mathbf{x})$. This is in contrast to computing many FFT's and inverse FFT's, each with a cost of $\mathcal{O}(G^d \log G)$, to solve the matrix eigenproblem. In essence, we can choose M larger than G with no significant additional computational cost, up to the point where the setup cost is approximately equal to the cost of solving the matrix eigensystem. Another factor that inhibits us from choosing very large M is the memory requirement for the storage of a $M^d \times M^d$ matrix of Fourier coefficients/nodal values.

In conclusion, approximating the Fourier coefficients of $\gamma(\mathbf{x})$ appears to be a significant handicap because of the large errors that are introduced. To alleviate this using the method we have described, we should choose M larger than G, but we do not yet know how much larger we should choose M. Depending on what our numerical observations tell us about our strategy for choosing M as a function of G we may obtain a method where the cost of computing the approximate Fourier coefficients of $\gamma(\mathbf{x})$ exceeds the cost of solving the matrix eigenproblem from our method.

We now present some results from numerical experiments to support our theory.

4.4.3 Examples

In this subsection we apply the sampling method to Model Problems 1-4 to support our theoretical error bounds for the sampling method. In the following plots, the reference solution is the solution to Problem 4.17 with $G = 2^{18} - 1$ for Model Problems 1 and 2 and $G = 2^{10} - 1$ for Model Problems 3 and 4. All of the following plots have logarithmically scaled axes.

In Figures 4-24 - 4-26 we plot the errors from the sampling method for fixed G and varying M.

In Figure 4-24 we plot the errors for Model Problem 1 and Model Problem 1a where Model Problem 1a is the same as Model Problem 1 except we have changed the ratio of glass to air in the photonic crystal from 50:50 to 55:45. We have introduced Model Problem 1a because Model Problem 1 appears to be a special case for the sampling method. For Model Problem 1 we observe that the eigenvalue errors are $\mathcal{O}(M^{-2})$, whereas for Model Problem 1a they are only $\mathcal{O}(M^{-1})$. We also observe that the eigenfunction errors of Model Problem 1 decay slightly quicker than $\mathcal{O}(M^{-1})$ while Model Problem 1a clearly exhibits $\mathcal{O}(M^{-1})$ decay. The observation that Model Problem 1 is a special case is reinforced when we consider the convergence rates of Model Problems 2-4.

In Figures 4-25 and 4-26 we observe that both the eigenvalue and eigenfunction errors of Model Problems 2-4 are $\mathcal{O}(M^{-1})$. This shows that the bounds that we proved in Theorem 4.42 are not sharp, and they should be $\mathcal{O}(M^{-1})$ instead of $\mathcal{O}(M^{-1/2})$.

With this observed error dependence on M we now optimise the errors by choosing $M = CG^r$ for a constants C and r. Our aim is to recover the convergence rates of

the spectral Galerkin method without sampling with the smallest amount of additional computational effort, i.e. we want to recover $\mathcal{O}(G^{-3/2})$ for the eigenfunction errors and $\mathcal{O}(G^{-3})$ for the eigenvalue errors with the smallest possible M. A simple calculation (using the observation that the eigenvalue and eigenfunction errors are $\mathcal{O}(M^{-1})$ rather than the bound in Theorem 4.42) shows that the eigenfunction convergence rate is recovered provided that we choose $M \geq G^{3/2}$ and the eigenvalue convergence rate is recovered if we choose $M \geq G^3$. For implementation, we should ensure that $M = 2^n$ for some $n \in \mathbb{N}$ (for best FFT performance). Therefore, we set $M = N_f^r$. This corresponds to choosing a constant $C \neq 1$ in $M = CG^r$. To minimise the additional computational cost we should choose $M = G^{3/2}$ for the eigenfunctions and $M = G^3$ for the eigenvalues.

In practice, with $M = G^{3/2}$ the setup cost is approximately the same as the cost of solving the matrix eigenproblem, but with $M = G^3$ we either get a method where the setup cost exceeds the cost of solving the matrix eigenproblem or we run out of computer memory for storing the $M^2 \times M^2$ matrix of sampled $\gamma(\mathbf{x})$ values. Therefore, in the case of the eigenvalue errors, the sampling method adds a significant amount of error that can not always be avoided.

We will now experiment with different strategies for choosing $M = N_f^r$ with different constants r to demonstrate that our error optimisation strategy is correct.

First, we consider the eigenfunction errors. In Figures 4-27 and 4-28 we plot the 1st eigenfunction errors of Model Problems 1a and 2-4 (since Model Problem 1 was a special case) for $r = 1, \frac{3}{2}, 2$. We observe that we achieve errors that are $\mathcal{O}(G^{-3/2})$ (same as standard method with exact Fourier coefficients) when $r = \frac{3}{2}$ and $r = 2$, but we only get $\mathcal{O}(G^{-1})$ errors when $r = 1$. Since there is more computational effort required when $r = 2$, this confirms that $r = \frac{3}{2}$ is the best strategy to minimise the eigenfunction errors with the least amount of extra computational work. Unsurprisingly, we do not observe errors that are smaller than the errors for the standard method for any choice of r.

Now we consider the strategy for choosing r to minimise the eigenvalue errors. In Figures 4-29 - 4-31 we plot the 1st eigenvalue errors of Model Problems 1a and 2-4 for different choices of r. We see (most clearly in Figure 4-29 for Model Problem 1a) that the we recover $\mathcal{O}(G^{-3})$ convergence when $M = N_f^3$. Unfortunately, memory constraints have limited our ability to compute many points for this case in all of the model problem examples.

In conclusion, it is possible to recover the convergence rates for the eigenvalues and eigenfunctions that we saw for the standard method by choosing M wisely. However, to achieve this there is a significant amount of extra computational work required (especially for eigenvalue calculations), and in some cases this extra work is prohibitively expensive. In these cases we must choose M as large as practicable and the errors will be dominated by the sampling method error.

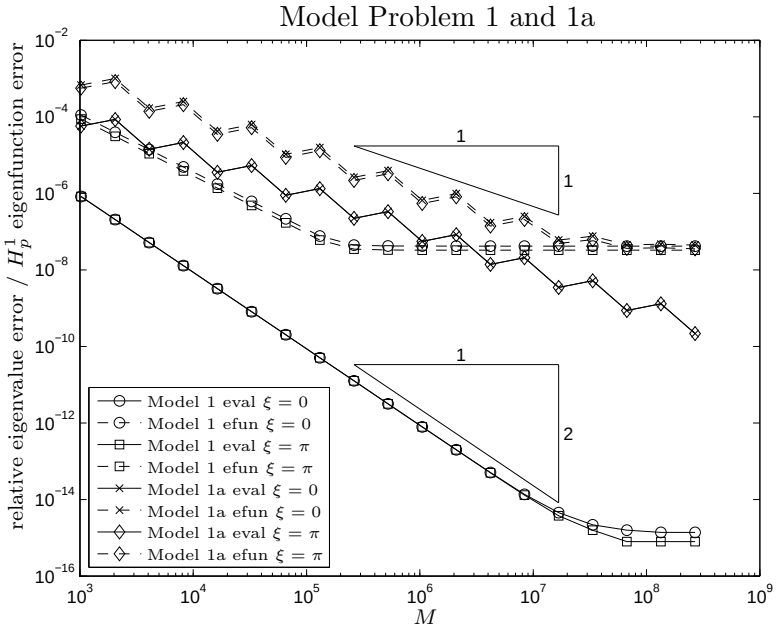

Figure 4-24: Plot of the error vs. M for Problem 4.40 (fixed G) for Model Problem 1 and 1a. The reference solution is the solution to Problem 4.17 with $G = 2^{18} - 1$.

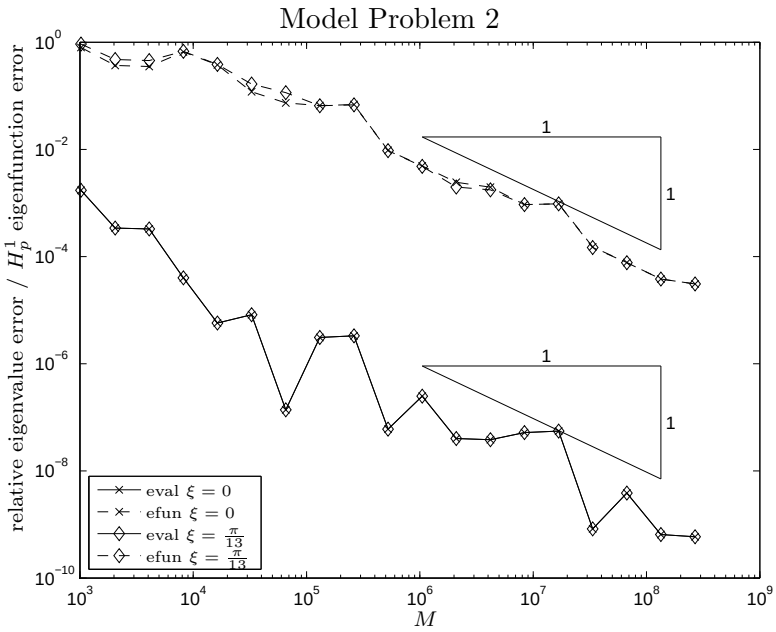

Figure 4-25: Plot of the error vs. M for Problem 4.40 (fixed G). The reference solution is the solution to Problem 4.17 with $G = 2^{18} - 1$.

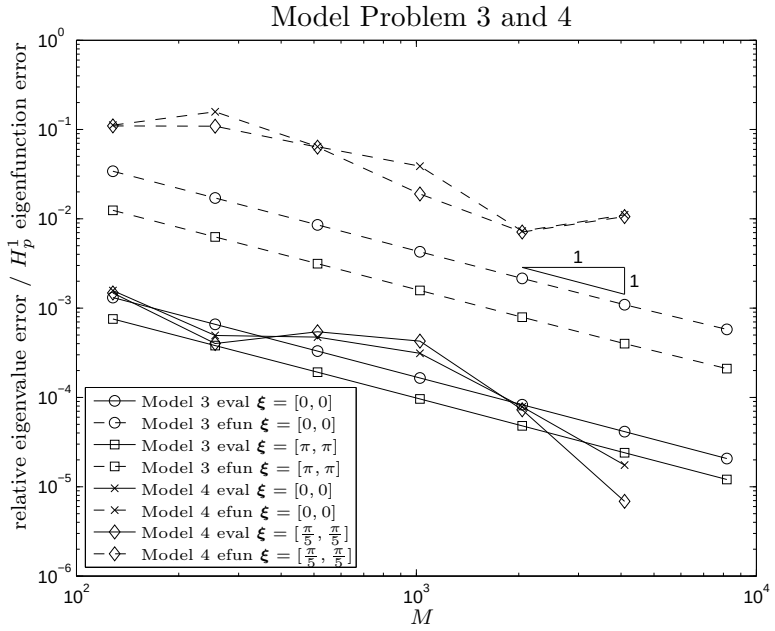

Figure 4-26: Plot of the error vs. M for Problem 4.40 (fixed G). The reference solution is the solution to Problem 4.17 with $G = 2^{18} - 1$.

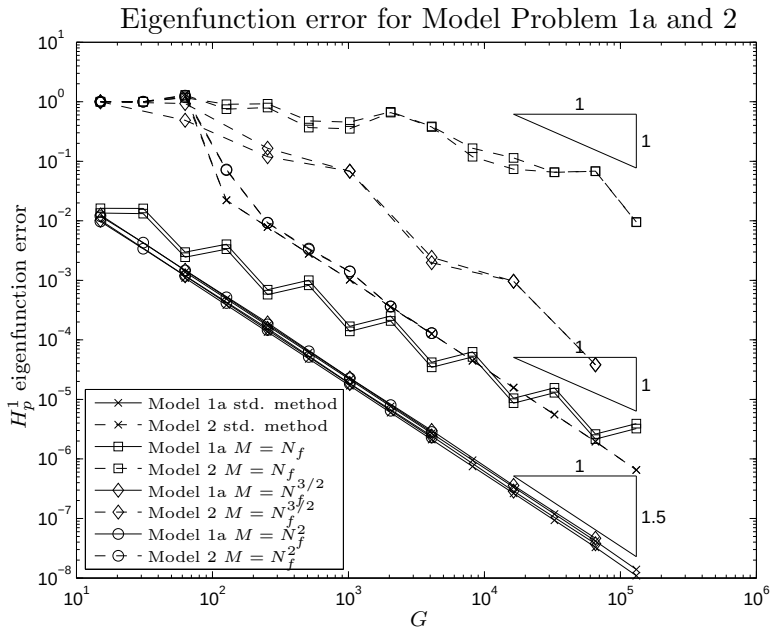

Figure 4-27: Plot of the 1st eigenfunction error vs. G for Problem 4.40. The reference solution is Problem 4.17 with $G = 2^{18} - 1$.

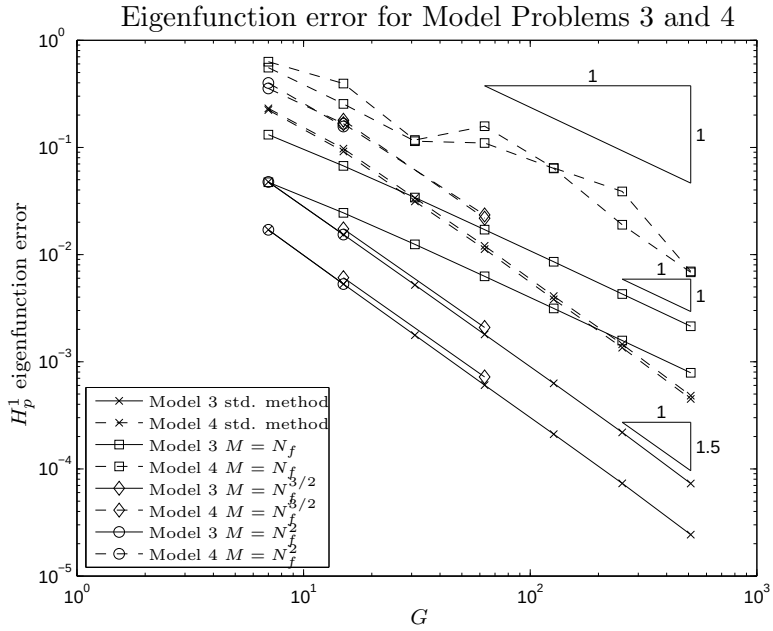

Figure 4-28: Plot of the 1st eigenfunction error vs. G for Problem 4.40. The reference solution is Problem 4.17 with $G = 2^{18} - 1$.

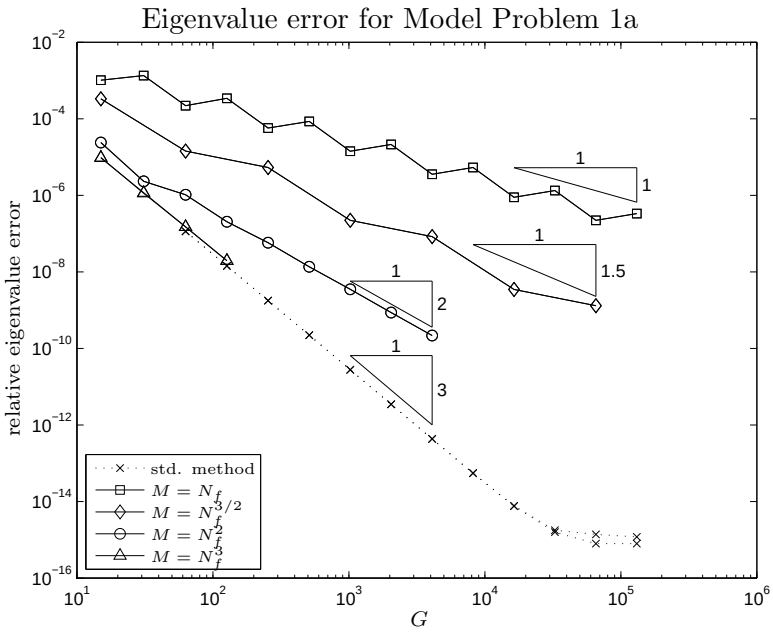

Figure 4-29: Plot of the 1st eigenvalue error vs. G for Problem 4.40. The reference solution is Problem 4.17 with $G = 2^{18} - 1$.

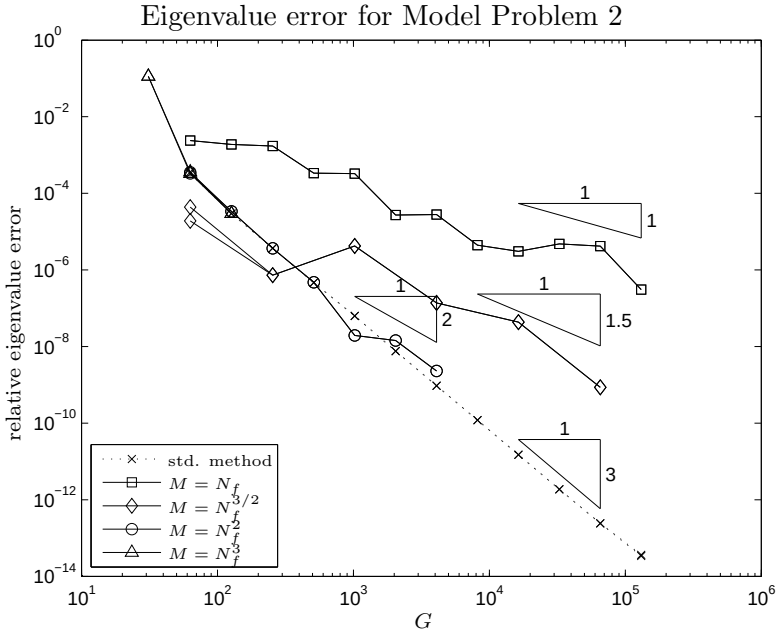

Figure 4-30: Plot of the 1st eigenvalue error vs. G for Problem 4.40. The reference solution is Problem 4.17 with $G = 2^{18} - 1$.

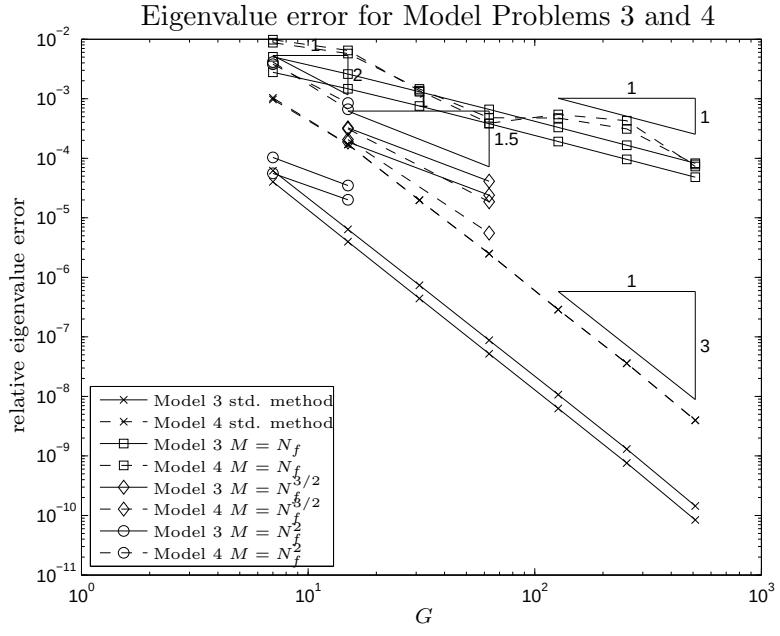

Figure 4-31: Plot of the 1st eigenvalue error vs. G for Problem 4.40. The reference solution is Problem 4.17 with $G = 2^{18} - 1$.

4.5 Smoothing and Sampling

In the final section of this chapter we put together our analysis of the smoothing method and the sampling method to analyse a method that uses both of these techniques simultaneously, as in [64].

In the previous section we saw that the sampling method provides us with an efficient method for approximating the Fourier coefficients of $\gamma(\mathbf{x})$. However, there was an additional error that was particularly significant for the eigenvalues. It is thought that this new method, that uses smoothing and sampling, will have smaller errors than the sampling method, or it might allow "rough" calculations to be made with relatively few plane waves. Both our analysis and numerical experiments will show that this new method does not yield faster convergence or smaller errors. However, our observations are inconclusive as to whether or not "rough" calculations are possible with smoothing and sampling instead of just sampling and this could be an area for further investigation.

The section is divided into three subsections. In the first subsection we describe the method, in the second subsection we perform the error analysis, and in the third subsection we present some numerical examples.

We assume that $\gamma(\mathbf{x}) \in PC'_p$ throughout this section.

4.5.1 The Method

The method for smoothing and sampling is the same as for the sampling method (Subsection 4.4.1), except we replace $[\mathrm{Q}_M \gamma]_{\mathbf{g}}$ with $[\widetilde{\mathrm{Q}_M \gamma}]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$, where $\widetilde{\mathrm{Q}_M \gamma}$ denotes the Gaussian smoothed $\mathrm{Q}_M \gamma$ and we defined Gaussian smoothing in Section 4.3.

To compute $[\widetilde{\mathrm{Q}_M \gamma}]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$ we first use Algorithm 4.36 to compute $[\mathrm{Q}_M \gamma]_{\mathbf{g}}$ for $\mathbf{g} \in \mathbb{Z}^2_{N_f,\square}$. Then we use the formula in Part 1 of Lemma 4.26 to get

$$[\widetilde{\mathrm{Q}_M \gamma}]_{\mathbf{g}} = \mathrm{e}^{-2\pi^2|\mathbf{g}|^2\Delta^2}[\mathrm{Q}_M \gamma]_{\mathbf{g}} \qquad \text{for all } \mathbf{g} \in \mathbb{Z}^2_{N_f,\square}.$$

The $[\widetilde{\mathrm{Q}_M \gamma}]_{\mathbf{g}}$ are then used instead of $[\gamma]_{\mathbf{g}}$ in Algorithm 4.19. Thus, the cost for computing the smoothing and sampling method has the same order as the cost for computing the sampling method and the memory requirements are the same.

Note that the smoothing we have applied acts as a filter *after* sampling.

4.5.2 Error Analysis

As we saw for the sampling method the error convergence rates for this method will depend on how $\|\gamma - \widetilde{\mathrm{Q}_M \gamma}\|_{H_p^{-1}}$ behaves with respect to Δ and M (recall that Δ determines the amount of Gaussian smoothing). Here we present a relatively simple

proof for a result that says: smoothing and sampling is at least as good as the sampling method, provided Δ is chosen appropriately. It does not show that smoothing and sampling is in any way better than sampling.

Lemma 4.43. *Let $d = 1, 2$, $\gamma \in PC'_p$ and define $\widetilde{\mathrm{Q}_M \gamma} := \mathcal{G} * \mathrm{Q}_M \gamma$ as in Subsection 3.2.5 and (4.20). With $-1 \leq s \leq 0$ and $\epsilon > 0$ we get:*

$$\|\gamma - \widetilde{\mathrm{Q}_M \gamma}\|_{H_p^s} \lesssim \Delta^{-s+1/2} + M^{-1/2+\epsilon}. \tag{4.50}$$

Proof. With $\widetilde{\gamma} = \mathcal{G} * \gamma$ we split $\|\gamma - \widetilde{\mathrm{Q}_M \gamma}\|_{H_p^s}$ into two parts

$$\|\gamma - \widetilde{\mathrm{Q}_M \gamma}\|_{H_p^s} \leq \underbrace{\|\gamma - \widetilde{\gamma}\|_{H_p^s}}_{I_1} + \underbrace{\|\widetilde{\gamma} - \widetilde{\mathrm{Q}_M \gamma}\|_{H_p^s}}_{I_2}.$$

From Part 2 of Lemma 4.26 we get $I_1 \lesssim \Delta^{-s+1/2}$. For I_2 we realise that $\widetilde{\gamma} - \widetilde{\mathrm{Q}_M \gamma} = \mathcal{G} * (\gamma - \mathrm{Q}_M \gamma)$. Part 1 of Lemma 4.26 then tells us that

$$\left| \left[\widetilde{\gamma} - \widetilde{\mathrm{Q}_M \gamma} \right]_{\mathbf{g}} \right| = \mathrm{e}^{-2\pi^2 |\mathbf{g}|^2 \Delta^2} \left| [\gamma - \mathrm{Q}_M \gamma]_{\mathbf{g}} \right| \leq \left| [\gamma - \mathrm{Q}_M \gamma]_{\mathbf{g}} \right| \tag{4.51}$$

for all $\mathbf{g} \in \mathbb{Z}^2$. Therefore, we get $\|\widetilde{\gamma} - \widetilde{\mathrm{Q}_M \gamma}\|_{H_p^s} \leq \|\gamma - \mathrm{Q}_M \gamma\|_{H_p^s} \leq M^{-1/2+\epsilon}$ using Lemma 4.39. □

In (4.51) it might appear as though we are being too convservative in throwing away the exponential term but the $\mathbf{g} = 0$ case is sharp.

Now we use exactly the same approach as for the error analysis of the sampling method in Subsection 4.4.2. First we define the discrete variational eigenvalue problem that our smoothing and sampling method is actually solving.

Problem 4.44. Find $\lambda_G \in \mathbb{R}$ and $0 \neq u_G \in \mathcal{S}_G$ such that

$$a_{\widetilde{\mathrm{Q}}}(u_G, v_G) = \lambda_G b(u_G, v_G) \qquad \forall v_G \in \mathcal{S}_G \tag{4.52}$$

where

$$a_{\widetilde{\mathrm{Q}}}(u, v) = \int_\Omega (\nabla + i\xi)\, u \cdot \overline{(\nabla + i\xi)\, v} + \left(\mathrm{K} - \widetilde{\mathrm{Q}_M \gamma} \right) u \overline{v} dx.$$

Now, we quote the main result, with the proof being the same as in Subsection 4.4.2.

Theorem 4.45. *Let λ be an eigenvalue of Problem 4.6 (with $\gamma \in PC'_p$) with multiplicity m and corresponding eigenspace $\mathcal{M}$. Then for sufficiently large G, large M and small $\Delta > 0$ there exist m eigenvalues $\lambda_1(G, \Delta, M), \ldots, \lambda_m(G, \Delta, M)$ of Problem 4.44 with*

corresponding eigenspaces $M_1(\lambda_1), \ldots, M_m(\lambda_m)$ *and a space*

$$\mathcal{M}_{G,\Delta,M} := \bigoplus_{j=1}^{m} M_j(\lambda_j)$$

such that for $\epsilon > 0$*,*

$$\delta(\mathcal{M}, \mathcal{M}_{G,\Delta,M}) \lesssim G^{-3/2+\epsilon} + \Delta^{3/2} + M^{-1/2+\epsilon}$$

and

$$|\lambda - \lambda_j| \lesssim G^{-3+2\epsilon} + \Delta^{3/2} + M^{-1/2+\epsilon} \qquad \textit{for } j = 1, \ldots, m.$$

From the numerical results in the previous sections we do not expect that the bounds in Theorem 4.45 are sharp. Instead, we expect that the eigenfunction error bound in Theorem 4.45 should be

$$\delta(\mathcal{M}, \mathcal{M}_{G,\Delta,M}) \lesssim G^{-3/2+\epsilon} + \Delta^{3/2} + M^{-1+\epsilon}$$

and the eigenvalue error bound should have the form

$$|\lambda - \lambda_j| \lesssim G^{-3+2\epsilon} + \Delta^{2} + M^{-1+\epsilon} \qquad \text{for } j = 1, \ldots, m.$$

Let us now consider some numerical examples to decide how to balance the error contributions by choosing Δ and M as functions that depend on G.

4.5.3 Examples

In this subsection we apply the smoothing and sampling method to Model Problems 1a, 2 and 3 to support our theoretical error bounds for the smoothing and sampling method. We calculate the error of Problem 4.44 for varying G where we have chosen $\Delta = G^{-r}$ and $M = N_f^s$ for different constants r and s. As a benchmark, we also plot the errors of the standard method which uses exact Fourier coefficients of $\gamma(\mathbf{x})$.

In the previous sections we saw that the smoothing method and the sampling method could not improve the convergence rate of the standard method. Here, we also expect this to be the case, but we will be interested in strategies for choosing the smoothing and sampling that recover the performance of the standard method.

We do not consider Model Problem 1 because it was a special case for the sampling method, and we do not plot any results for Model Problem 4 because the errors have not entered the asymptotic regime for the range of G that we consider and choosing larger G is beyond the memory capabilities of the computer we used for the computations.

In the following plots, the reference solution is the solution to Problem 4.17 with $G = 2^{18} - 1$ for Model Problems 1a and 2 and $G = 2^{10} - 1$ for Model Problem 3.

We first consider the eigenfunction errors of our model problems, which are plotted in Figures 4-32 - 4-34. For all of these plots we see that the fastest rate of decay is $\mathcal{O}(G^{-3/2})$, as for the standard method. Moreover, the $\mathcal{O}(G^{-3/2})$ rate of decay is only achieved when $s = \frac{3}{2}$. Therefore, we recommend the strategy of choosing $M = N_f^{3/2}$. This strategy is the same as for the sampling method without smoothing. It appears that our strategy for choosing $s = \frac{3}{2}$ is independent from our choice of r for the values of r that we have plotted.

From the plots it appears that the best strategy for choosing r is $r = \frac{3}{2}$ (or $r = 2$ for Model Problem 3), which corresponds to smaller Δ and less smoothing. Ultimately, we observe that less smoothing is better and we therefore recommend choosing $\Delta = 0$ and reverting back to the sampling method. However, since the optimal rate of decay is also achieved for $r = 1$, and $r > 1$ corresponds to smaller Δ, we could potentially recover the performance of the standard method by choosing any $\Delta \leq CG^{-r}$ with $r = 1$ and a fixed constant $C \ll 1$.

Now let us consider the eigenvalue errors of our model problems. These are plotted in Figures 4-35 - 4-37. For Model Problem 1a in Figure 4-35 we see that we should choose s as large as possible ($s = 2$ is the largest that we have plotted) and $r \geq \frac{3}{2}$ to achieve the best results, but unlike the eigenfunction errors we do not recover the convergence rate of the standard method. Choosing $s = 2$ corresponds to choosing $M = N_f^2$ which is the largest M that we can compute with. Perhaps if we could do computations for $s = 3$ we would recover $\mathcal{O}(G^{-3})$ convergence, but we are limited to $s = 2$ by computer memory restraints. Choosing $r = \frac{3}{2}$ corresponds to the largest amount of smoothing that is permissible without adding a significant error. Therefore, choosing any $r \geq \frac{3}{2}$ is an acceptable strategy that will recover the optimal convergence rate, $\mathcal{O}(G^{-3})$. In fact, we could choose $\Delta = 0$ without penalty and revert to the sampling method.

The eigenvalue error plots for Model Problems 2 and 3 are not as clean as the plot for Model Problem 1a but we can still see the overall theme: we get the smallest errors when M is as large as practicable and when Δ is sufficiently small. Moreover, we do not see errors decay at a rate that is faster than the optimal rate, $\mathcal{O}(G^{-3})$.

In conclusion, we have not found any evidence that smoothing *with* sampling is in any way a better method than the sampling method *without smoothing*. Indeed, when we have been optimising our choice of smoothing by choosing r we have essentially been ensuring that the smoothing is sufficiently small as to not contribute to the overall error. It still remains open as to whether or not smoothing will assist in making "rough" calculations and this requires further investigation.

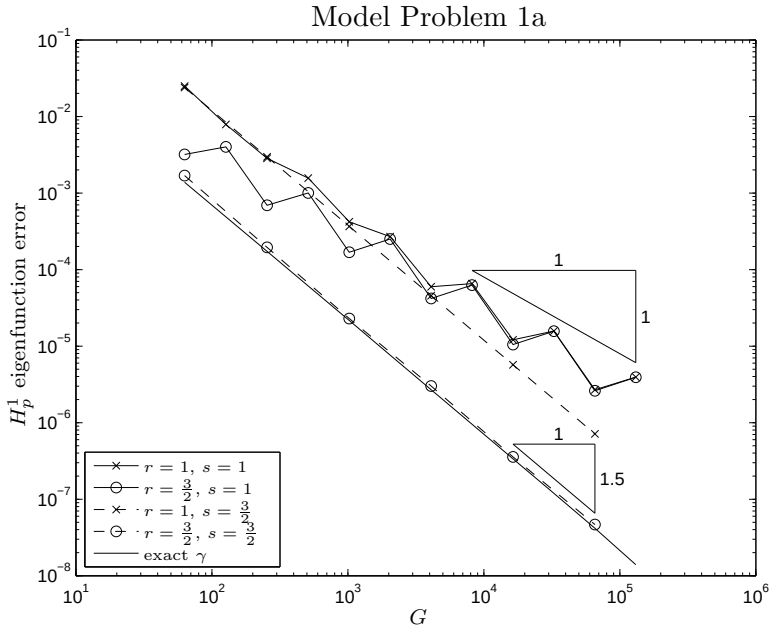

Figure 4-32: Plot of the 1st eigenfunction error vs. G for Problem 4.40 where we have chosen $\Delta = G^{-r}$ and $M = N_f^s$ for different constants r and s.

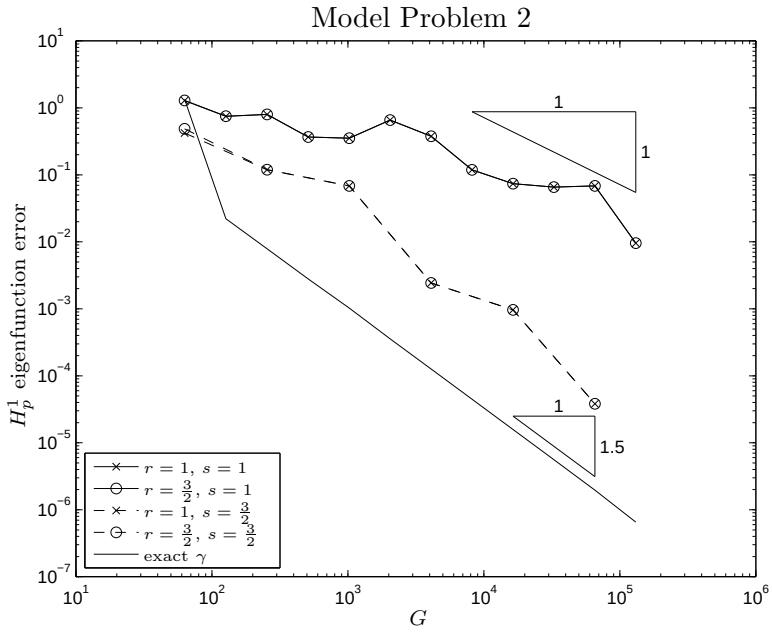

Figure 4-33: Plot of the 1st eigenfunction error vs. G for Problem 4.40 where we have chosen $\Delta = G^{-r}$ and $M = N_f^s$ for different constants r and s.

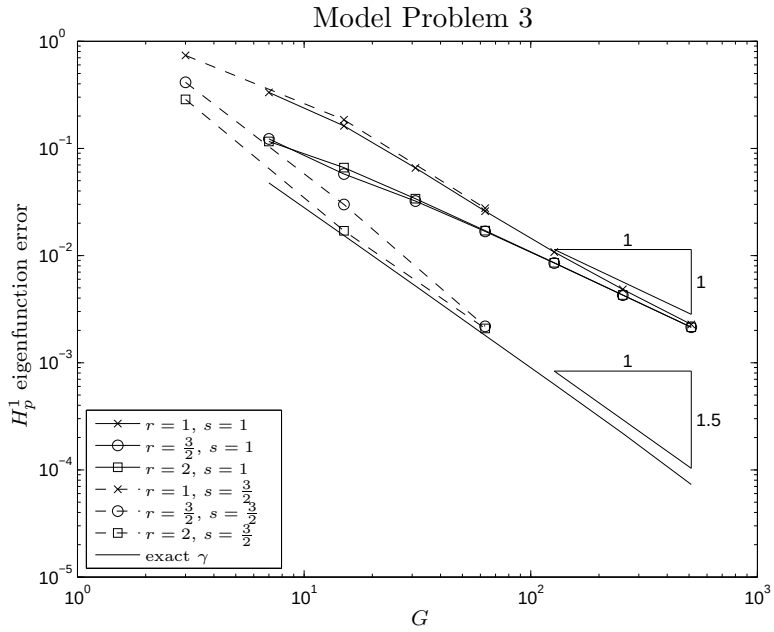

Figure 4-34: Plot of the 1st eigenfunction error vs. G for Problem 4.40 where we have chosen $\Delta = G^{-r}$ and $M = N_f^s$ for different constants r and s. The reference solution is Problem 4.17 with $G = 2^{18} - 1$.

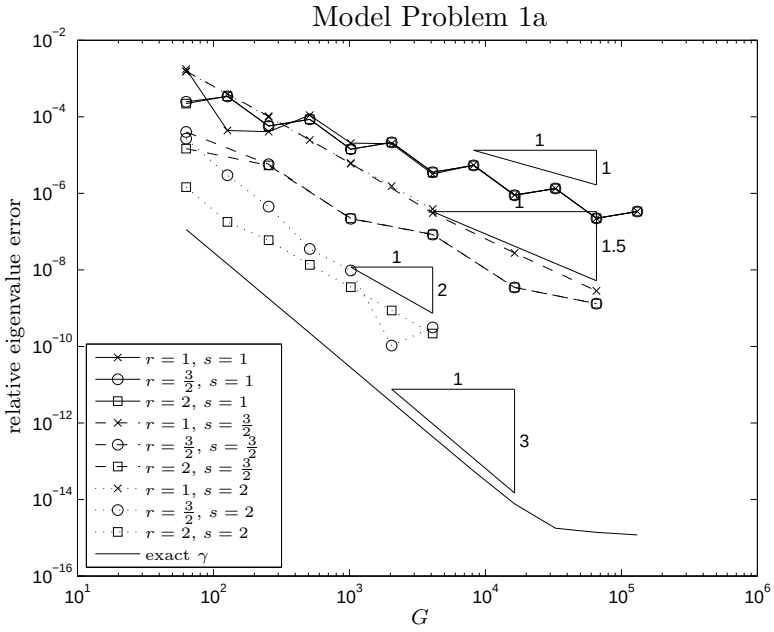

Figure 4-35: Plot of the 1st eigenvalue error vs. G for Problem 4.40 where we have chosen $\Delta = G^{-r}$ and $M = N_f^s$ for different constants r and s. The reference solution is Problem 4.17 with $G = 2^{18} - 1$.

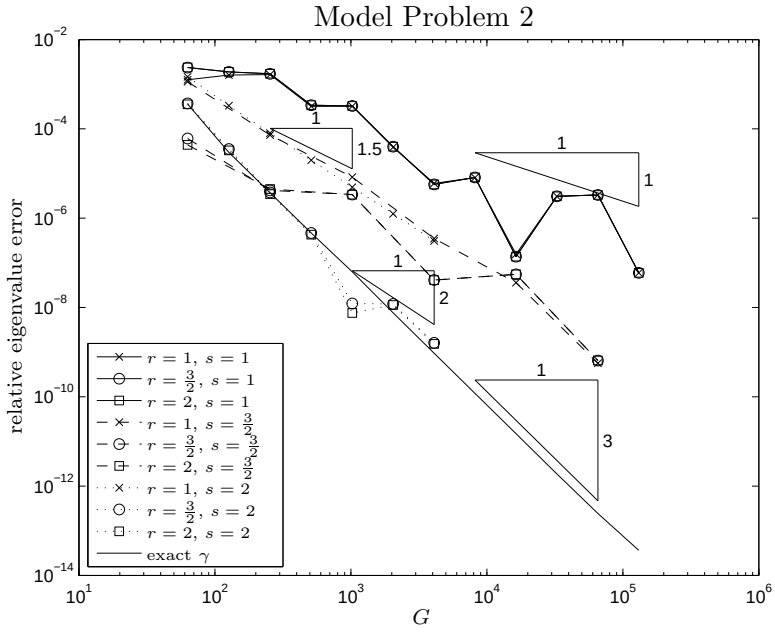

Figure 4-36: Plot of the 1st eigenvalue error vs. G for Problem 4.40 where we have chosen $\Delta = G^{-r}$ and $M = N_f^s$ for different constants r and s.

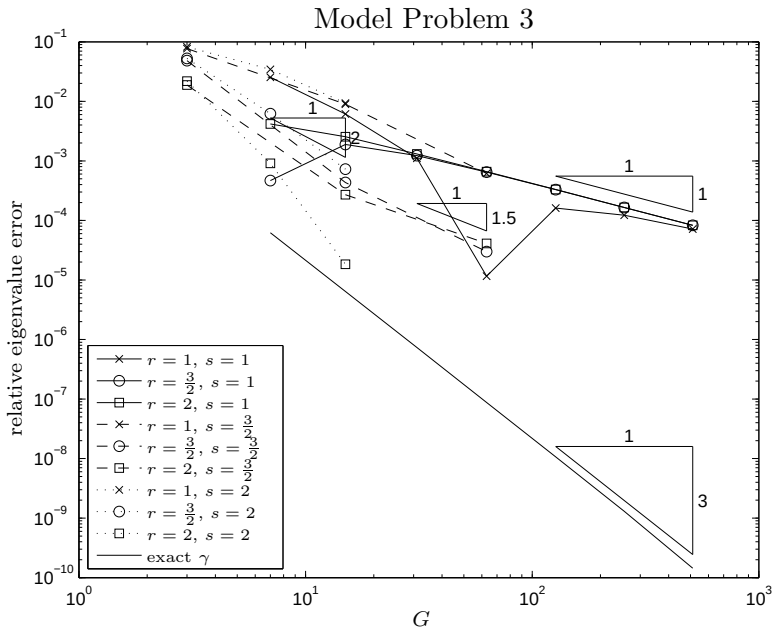

Figure 4-37: Plot of the 1st eigenvalue error vs. G for Problem 4.40 where we have chosen $\Delta = G^{-r}$ and $M = N_f^s$ for different constants r and s. The reference solution is Problem 4.17 with $G = 2^{18} - 1$.

4.6 Curvilinear Coordinates

Finally, and briefly, we make a remark about another variation of the plane wave expansion method that has been used in [63] and [64]. In this method $\gamma(\mathbf{x})$ and $u(\mathbf{x})$ are sampled on a non-uniform grid, unlike in Section 4.4. This is intended to allow the sampling nodes to be more concentrated near the material interfaces and therefore provide a better approximation of $\gamma(\mathbf{x})$ and of $u(\mathbf{x})$. In [63] and [64] the method is presented and the author cleverly devises a way of computing matrix-vector products with the system matrix whilst preserving the efficiency ($\mathcal{O}(N \log N)$ operations), albeit with 6 FFTs instead of the 2 FFTs that are currently required. The additional FFTs arise because the Laplacian part of the matrix is no longer confined to the diagonal (c.f. (4.14)). This is because the expansion terms are no longer orthogonal. An important consequence of the Laplacian part of the matrix no longer being confined to the diagonal is that the simple preconditioners ((4.18) and (4.19)) no longer "cancel" the Laplacian part of the operator and are no longer optimal. A method for obtaining an optimal preconditioner to use with a curvilinear coordinate expansion method would require further investigation and we do not consider this method any further in this thesis. We only mention that without a suitable preconditioner this method very quickly becomes very costly to compute and thus unfeasible. It is also not immediately obvious in what way the curvilinear expansion improves the approximation error for a fixed number of expansion terms, and how one would go about proving an improved error bounds with a faster convergence rate.

CHAPTER 5

1D TM MODE PROBLEM

In this chapter we consider the errors from the plane wave expansion method applied to the 1D TM Mode Problem, Problem 2.4 (in Section 2.5). The error analysis is not as straight forward as for the Scalar 2D Problem and for the 1D TE Mode Problem in Chapter 4.

We begin by applying results from [25] to obtain a variational eigenvalue problem to solve. To do this we consider the 1D TM Mode Problem written in *divergence form*, (2.22), and we quote some results from [25]. We then present the implementation details for the plane wave expansion method applied to this problem. We do this by following the technique used in [64] and [39] where plane wave expansions of the eigenfunction and coefficient functions are substituted into the governing equation before neglecting high-frequency terms to get a finite dimensional problem. This is in contrast to how we presented the plane wave expansion method in Chapter 4 where we presented it as a Galerkin method.

To begin the error analysis we develop regularity results for the variational eigenproblem that corresponds to the divergence form of the 1D TM Mode Problem. We see that the 1D TM Mode Problem has less regularity than the 1D TE Mode Problem. We then develop error analysis for the spectral Galerkin method applied to this problem using the same techniques that we used in Chapter 4 for the 1D TE Mode Problem and the Scalar 2D Problem. Unfortunately, this method is not equivalent to the plane wave expansion method and it can not be implemented as efficiently as the plane wave expansion method. To develop error analysis for the plane wave expansion method we write the method in terms of the variational eigenproblem corresponding to the divergence form of the 1D TM Mode Problem and we discover that it is equivalent to a non-conforming Petrov-Galerkin method. Unfortunately, using the existing theory for Petrov-Galerkin methods does not yield the required results. Nevertheless, it still

seems to be the most promising route for future investigations. We can, however, derive approximation error results for the exact eigenfunctions approximated with plane waves using the regularity results that we developed earlier. These approximation error results give us an upper limit for the rate at which the plane wave expansion method can converge (to the exact eigenfunctions). With numerical examples we then confirm that the plane wave expansion method actually achieves this upper limit and it converges at the fastest possible rate, given the limited regularity of the exact eigenfunctions. Note that this is substantially lower than for the 1D TE Mode Problem, namely $\mathcal{O}(N^{-1/2})$ instead of $\mathcal{O}(N^{-3/2})$ where N is the number of plane waves.

As well as computing numerical examples for the standard plane wave expansion method we also present numerical examples for smoothing and sampling within the plane wave expansion method. As for the 1D TE Mode Problem we observe that smoothing does not improve the convergence of the plane wave expansion method, and the sampling method requires a sufficiently fine sampling grid to recover the convergence rate of the standard plane wave expansion method (where exact Fourier coefficients are used).

The main motivation for studying the 1D TM Mode Problem is to gain insight into the behaviour of the Full 2D Problem since the 1D TM Mode Problem can be thought of as a restriction to 1D of the the Full 2D Problem.

5.1 The Problem

Formally, the 1D TM mode problem (see Problem 2.4) is

$$\frac{d^2h}{dx^2} + \gamma(x)h - \frac{d\eta}{dx}\frac{dh}{dx} = \beta^2 h \tag{5.1}$$

where $h = h(x)$ is an eigenfunction, $\gamma(x) = \frac{4\pi^2 n^2(x)}{\lambda_0^2}$ and $\eta(x) = \log n^2(x)$ are periodic and piecewise constant, and β^2 is an eigenvalue. More details about this equation are given in Chapter 2. As in Chapter 4 we restrict $n^2(x)$ so that it is periodic with period cell $\Omega = [-\frac{1}{2}, \frac{1}{2}]$ and $1 \leq n^2(x) \leq n_{max}^2$. More specifically, we assume that $n^2(x)$ is discontinous at points $x_j \in \Omega$ for $j = 1, \ldots, J$. We then divide Ω into intervals $\Omega_j = (x_j, x_{j+1})$ for $j = 1, \ldots, J$ (we define $x_{J+1} := x_1 + 1$) and specify that $n^2(x) = n_j^2$ for $x \in \Omega_j$ where the n_j are constants. For notational purposes let us define $x_{j+\frac{1}{2}}$ as the midpoint of the interval Ω_j, i.e. define $x_{j+\frac{1}{2}} := \frac{1}{2}(x_j + x_{j+1})$ for $j = 1, \ldots, J$.

This problem can be rewritten in divergence form (see (2.22)),

$$\frac{d}{dx}\left(\frac{1}{n^2}\frac{dh}{dx}\right) + ch = \frac{\beta^2}{n^2}h \tag{5.2}$$

where $c = \frac{4\pi^2}{\lambda_0^2}$ is constant.

Applying Floquet/Bloch theory to this problem is not as straight forward as for the 1D TE Mode Problem or the Scalar 2D Problem in the previous chapter. However, this issue has been addressed in [25]. According to [25] there exists a linear non-negative self-adjoint operator on a Hilbert space that corresponds to (5.2) (whose action is expressed in terms of a quadratic form). Moreover, Floquet/Bloch theory can be applied (through the quadratic forms) to obtain a family of problems to solve, from which we can recover the spectrum of the original operator. Each member of the new family of problems is given below.

Problem 5.1. For $\xi \in B := [-\pi, \pi]$, find $\lambda \in \mathbb{C}$ and $0 \neq u \in H_p^1$ such that

$$a(u, v) = \lambda b(u, v) \qquad \forall v \in H_p^1$$

where

$$a(u,v) = \int_\Omega \tfrac{1}{n^2}\left((\tfrac{d}{dx} + i\xi)u\overline{\left(\tfrac{d}{dx} + i\xi\right)v} + (\mathrm{K} - cn^2)u\overline{v}\right)dx$$
$$b(u,v) = \int_\Omega \tfrac{1}{n^2}u\overline{v}dx$$

and $\mathrm{K} \geq cn_{max}^2 + 2\pi^2 n_{max}^4 + \frac{1}{2}$.

According to [25] there exists a non-negative self-adjoint operator on H_p^1 corresponding to this problem, and a result (Corollary 3.9 in [25]) that is equivalent to Theorem 3.63 also applies in this case from which we recover the spectrum of the original operator by solving Problem 5.1 for a range of $\xi \in B$.

We now restrict our attention to solving Problem 5.1 for fixed $\xi \in B$. For each $\xi \in B$, the bilinear form $a(\cdot, \cdot)$ is bounded and coercive.

Lemma 5.2. *The bilinear form $a(\cdot, \cdot)$ is bounded and coercive on H_p^1 provided we choose $\mathrm{K} \geq cn_{max}^2 + 2\pi^2 n_{max}^4 + \frac{1}{2}$.*

Proof. $a(\cdot, \cdot)$ bounded. Using a similar proof to the proof of Lemma 4.7 we get

$$\begin{aligned}
|a(u,v)| &= \left|\int_\Omega \tfrac{1}{n^2}\left((\tfrac{d}{dx} + i\xi)u\overline{\left(\tfrac{d}{dx} + i\xi\right)v} + (\mathrm{K} - cn^2)u\overline{v}\right)dx\right| \\
&\leq \|\tfrac{1}{n^2}\|_\infty \int_\Omega \left|(\tfrac{d}{dx} + i\xi)u\overline{\left(\tfrac{d}{dx} + i\xi\right)v} + (\mathrm{K} - cn^2)u\overline{v}\right| dx \\
&\leq \|\tfrac{1}{n^2}\|_\infty \left((1+\pi)^2 + \mathrm{K}\right)\|u\|_{H_p^1}\|v\|_{H_p^1} \qquad \forall u, v \in H_p^1.
\end{aligned}$$

$a(\cdot, \cdot)$ coercive. Using the Cauchy-Schwarz inequality and the arithmetic-geometric

mean inequality ($2ab \le a^2 + b^2$)

$$\begin{aligned}
a(v,v) &= \int_\Omega \tfrac{1}{n^2}\left(|(\tfrac{d}{dx} + i\xi)v|^2 + (\mathrm{K} - cn^2)|v|^2\right) dx \\
&= \int_\Omega \tfrac{1}{n^2}\left(|v'|^2 + i\xi v\overline{v}' - i\xi v'\overline{v} + |\xi|^2|v|^2 + (\mathrm{K} - cn^2)|v|^2\right) dx \\
&\ge \tfrac{1}{n_{max}^2}\left(|v|_{H^1}^2 + (|\xi|^2 + \mathrm{K} - cn_{max}^2)\|v\|_{L_p^2}^2\right) - 2|\xi||v|_{H^1}\|v\|_{L_p^2} \\
&= \tfrac{|v|_{H^1}^2}{n_{max}^2} + \tfrac{(|\xi|^2 + \mathrm{K} - cn_{max}^2)\|v\|_{L_p^2}^2}{n_{max}^2} - 2\left(\tfrac{|v|_{H^1}}{\sqrt{2n_{max}^2}}\right)\left(\sqrt{2n_{max}^2}|\xi|\|v\|_{L_p^2}\right) \\
&\ge \tfrac{1}{2n_{max}^2}|v|_{H^1}^2 + \left(\tfrac{\mathrm{K}}{n_{max}^2} - c - 2\pi^2 n_{max}^2\right)\|v\|_{L_p^2}^2 \\
&\ge \tfrac{1}{2n_{max}^2}\|v\|_{H_p^1}^2
\end{aligned}$$

provided we choose $\mathrm{K} \ge cn_{max}^2 + 2\pi^2 n_{max}^4 + \frac{1}{2}$. □

Following our approach from previous chapters we now define the solution operator $\mathrm{T} : L_p^2 \to H_p^1$ that corresponds to Problem **??**. As in Definition 3.70, for $f \in L_p^2$ we define $\mathrm{T}\,f \in H_p^1$ by

$$a(\mathrm{T}\,f, v) = b(f, v) \qquad \forall v \in H_p^1. \tag{5.3}$$

Theorem 5.3. *With* $\mathrm{T} : L_p^2 \to H_p^1$ *defined by* (5.3) *we get:*

1. $\mathrm{T} : H_p^1 \to H_p^1$ *is bounded, compact, positive definite and self-adjoint with respect to* $a(\cdot,\cdot)$.

2. $\sigma(\mathrm{T}) \subset \mathbb{R}$.

3. $\sigma(\mathrm{T})$ *is discrete, i.e.* $\sigma(\mathrm{T})$ *consists of nonzero isolated eigenvalues of finite multiplicity with no accumulation point.*

Proof. The proof for Part 1 is the same as the proof for Lemma 4.9. Parts 2 and 3 then follow from Theorem 3.60. □

By Lemma 3.71 we know that (μ, u) is an eigenpair of T if and only if $(\frac{1}{\mu}, u)$ is an eigenpair of the following variational eigenvalue problem. Note that $\mu \neq 0$ since T is positive.

Thus, the 1D TM Mode Problem can be solved by solving Problem 5.1. However, it is not yet clear how the plane wave expansion method can be expressed as a Galerkin method applied to Problem 5.1 so that we can apply the convergence theory in [6]. In the next section we present the details of the plane wave expansion method as we have implemented it, and we address the issue of how the plane wave expansion method relates to Problem 5.1 in Section 5.3.

5.2 Plane Wave Expansion Method and Implementation

In this section we present the plane wave expansion method applied to the 1D TM Mode Problem as it is presented in [64] or [39], and as we have implemented it. We do not apply the Galerkin method to Problem 5.1 because the $\frac{1}{n^2}$ factor in $a(\cdot,\cdot)$ ruins the orthogonality of the plane waves. This has the effect of causing the contributions from the derivatives in $a(\cdot,\cdot)$ to spill off the main diagonal of the matrix in the matrix eigenproblem (that is equivalent to the discrete variational problem we get from applying the Galerkin method). Another reason why we do not use the Galerkin method applied to Problem 5.1 is that we can not use the Fast Fourier Transform to efficiently compute matrix-vector products as we can for the method that we now present.

We begin by adding $\mathrm{K}\,h$ (where K is from the definition of Problem 5.1) to (5.1). Following the approach in [39] we can write $h(x) = u(x)\,\mathrm{e}^{i\xi x}$ for some $\xi \in B := [-\pi, \pi]$ where $u(x)$ is a periodic function. The equation we obtain is

$$-(\tfrac{d}{dx} + i\xi)^2 u + \tfrac{d(\log n^2)}{dx}(\tfrac{d}{dx} + i\xi)u - \gamma u + \mathrm{K}\,u = \lambda u \tag{5.4}$$

where $\gamma(x) = \frac{4\pi^2 n^2(x)}{\lambda_0^2}$ is the same as $\gamma(x)$ in Chapter 4. Thus, for each $\xi \in B$, we would like to solve (5.4) for eigenvalues λ and eigenfunctions h. In [3], it is claimed that solving the problem for all $\xi \in B$ is sufficient to obtain all possible eigenvalues and modes of (5.1). In [3] this is referred to as Bloch Theory.

To apply the plane wave expansion method to (5.4) we do the following: Expand u, γ and $\log n^2$ in terms of their plane wave expansions (or Fourier Series), for example,

$$u(x) = \sum_{g\in\mathbb{Z}} [u]_g\,\mathrm{e}^{i2\pi g x}\,.$$

Substitute the expansions of u, γ and $\log n^2$ into (5.4) to get

$$\sum_{g\in\mathbb{Z}} \left((\xi + 2\pi g)^2 - \sum_{k\in\mathbb{Z}} (2\pi k)[\log n^2]_k (\xi + 2\pi g)\,\mathrm{e}^{i2\pi k x} - \sum_{k\in\mathbb{Z}} [\gamma]_k\,\mathrm{e}^{i2\pi k x} + \mathrm{K} \right) [u]_g\,\mathrm{e}^{i2\pi g x} = \lambda \sum_{g\in\mathbb{Z}} [u]_g\,\mathrm{e}^{i2\pi g x} \tag{5.5}$$

Now multiply both sides of (5.5) by $\mathrm{e}^{i2\pi g' x}$ for $g' \in \mathbb{Z}$ and integrate over Ω to get

$$(\xi + 2\pi g')^2 [u]_{g'} - \sum_{g\in\mathbb{Z}} 2\pi(g' - g)[\log n^2]_{g'-g}(\xi + 2\pi g)[u]_g - \sum_{g\in\mathbb{Z}} [\gamma]_{g'-g}[u]_g + \mathrm{K}[u]_{g'} = \lambda [u]_{g'} \tag{5.6}$$

So far we have an infinite dimensional problem. To approximate h and λ and make the

problem finite dimensional we restrict g and g' so that $|g|, |g'| \leq G$ for a chosen $G \in \mathbb{N}$. Equivalently, we force $[u]_g = 0$ for all $|g| > G$ and we only consider (5.6) for $|g'| \leq G$. Equation (5.6) becomes

$$(\xi + 2\pi g')^2 [u]_{g'} - \sum_{|g| \leq G} 2\pi(g' - g)[\log n^2]_{g'-g}(\xi + 2\pi g)[u]_g - \sum_{|g| \leq G} [\gamma]_{g'-g}[u]_g + \mathrm{K}[u]_{g'} = \lambda_G [u]_{g'} \qquad \text{for } |g'| \leq G \quad (5.7)$$

The final step of the plane wave expansion method is to rewrite (5.7) as a $N \times N$ (where $N = 2G + 1$) matrix eigenvalue problem,

$$\mathrm{A}\,\mathbf{u} = \lambda_G \mathbf{u}, \tag{5.8}$$

where $\mathbf{u}$ is the N-vector with entries (by a slight abuse of notation) $u_g = [u]_g$ for $g = -G, \ldots, G$. The matrix A can be written as

$$\mathrm{A} = \mathrm{D} - \mathrm{W} - \mathrm{V}$$

where D is a diagonal matrix with diagonal entries $\mathrm{D}_{gg} = |\xi + 2\pi g|^2 + \mathrm{K}$, W is a full matrix with entries $\mathrm{W}_{gg'} = 2\pi(g - g')(\xi + 2\pi g')[\log n^2]_{g-g'}$, and V is the same matrix as in Section 4.2 with entries given by $\mathrm{V}_{gg'} = [\gamma]_{g-g'}$, for $g, g' = -G, \ldots, G$.

It remains to solve (5.8). We want to find the eigenvalues of (5.8) in the interval $[0, \mathrm{K}]$ and the corresponding eigenvectors (of which there are only finitely many, independent of G). We use the same implementation as in Subsection 4.2.2. However, this implementation again requires an efficient algorithm for computing matrix-vector products with A, and since A is non-symmetric, we use GMRES instead of PCG to obtain the action of A^{-1}. To compute $\mathrm{A}\,\mathbf{x}$ for a vector $\mathbf{x}$, we need to compute $\mathrm{D}\,\mathbf{x}$, $\mathrm{W}\,\mathbf{x}$ and $\mathrm{V}\,\mathbf{x}$. Computing $\mathrm{D}\,\mathbf{x}$ is easy because D is diagonal and we can compute $\mathrm{V}\,\mathbf{x}$ in $\mathcal{O}(N \log N)$ operations using the Fast Fourier Transform since V is Toeplitz. All we need now is an efficient algorithm to compute $\mathrm{W}\,\mathbf{x}$.

To compute $\mathrm{W}\,\mathbf{x}$ we first realise that we can write W as the product of two matrices,

$$\mathrm{W} = \mathrm{W}_1\,\mathrm{W}_2$$

where W_1 is Toeplitz and W_2 is diagonal, with entries

$$(\mathrm{W}_1)_{gg'} = 2\pi(g - g')[\log n^2]_{g-g'} \qquad \text{and} \qquad (\mathrm{W}_2)_{gg} = \xi + 2\pi(g)$$

for $g, g' = -G, \ldots, G$. Thus, to compute $\mathrm{W}\,\mathbf{x}$ we first compute $\mathbf{y} = \mathrm{W}_2\,\mathbf{x}$ in $\mathcal{O}(N)$ operations and then we compute $\mathrm{W}_1\,\mathbf{y}$ in $\mathcal{O}(N \log N)$ operations, again using the FFT. In summary, we see that we can compute $\mathrm{A}\,\mathbf{x} = (\mathrm{D} - \mathrm{W} - \mathrm{V})\mathbf{x}$ in $\mathcal{O}(N \log N)$ operations.

As well as using the FFT to efficiently compute matrix-vector products with A we also use a preconditioner to solve linear systems with A (to obtain the action of A^{-1}). For this problem we use exactly the same preconditioner as for the 1D TE Mode Problem, see (4.19), together with the GMRES algorithm to solve linear systems. We observe that this preconditioner is sufficient to guarantee that GMRES converges in $\mathcal{O}(1)$ iterations and that (provided K is sufficiently small) the Implicitly Restarted Arnoldi method solves (5.8) (for the fixed number of eigenpairs that we want) in $\mathcal{O}(1)$ iterations. Altogether, we can solve (5.8) in $\mathcal{O}(N \log N)$ operations.

5.3 Error Analysis

In this section we present the error analysis for two methods applied to Problem 5.1: The plane wave expansion method and the spectral Galerkin method. Unlike for the Scalar 2D Problem and the 1D TE Mode Problem, these two methods are not the same. We will find that the plane wave expansion method has implementation advantages but we can only do a full error analysis of the spectral Galerkin method.

We begin by proving a regularity result for Problem 5.1. In Chapter 4 we saw that the convergence properties of the plane wave expansion method were limited by the regularity of the eigenfunctions of the exact problem. Using the regularity result we also prove an approximation error estimate for eigenfunctions of Problem 5.1 approximated using plane waves.

Following the regularity result for Problem 5.1 we define the spectral Galerkin method and then investigate the convergence properties of this method. We consider this method before we consider the plane wave expansion method because we are able to use the same techniques that we used in Chapter 4 to analyse the error. Despite the ease with which we do a complete error analysis for the spectral Galerkin method, unfortunately, it does not share the same implementation efficiencies as the plane wave expansion method, as we discussed at the beginning of the previous section.

After our discussion of the spectral Galerkin method we return to the error analysis for the plane wave expansion method. First, we show that it is equivalent to two different variational problems: a Galerkin method where the bilinear form is not the same as that in Problem 5.1, and a non-conforming Petrov-Galerkin method applied to Problem 5.1. Neither of these presentations has so far lead to a complete error analysis and we have not been able to prove the stability of the plane wave expansion method. However, assuming stability of the method, we can nevertheless use the approximation error result for plane waves approximating eigenfunctions of Problem 5.1 to give us an upper limit for the rate of convergence of the plane wave expansion method. The numerical results in Section 5.4 suggest that such a stability result should be possible to prove.

5.3.1 Regularity

We start by proving a regularity result for eigenfunctions of Problem 5.1 and then use the regularity result to estimate the approximation error for plane wave approximation of the an eigenfunctions of Problem 5.1.

Theorem 5.4. *Let $f \in H_p^s$ for some $s \geq 0$. Define $f_j := f|_{\Omega_j}$ and $u_j := \mathrm{T}\, f|_{\Omega_j}$ for each $j = 1, \ldots, J$. Then*

1. *$u_j \in H^{s+2}(\Omega_j)$ and*
$$\|u_j\|_{H^{s+2}(\Omega_j)} \lesssim \|f_j\|_{H^s(\Omega_j)}$$

2. *$\frac{1}{n^2}(\frac{d}{dx} + i\xi)\,\mathrm{T}\, f \in H_p^1$ (and is therefore continuous by Theorem 3.27) and*
$$\|\tfrac{1}{n^2}\left(\tfrac{d}{dx} + i\xi\right)\mathrm{T}\, f\|_\infty \lesssim \|\tfrac{1}{n^2}\left(\tfrac{d}{dx} + i\xi\right)\mathrm{T}\, f\|_{H_p^1} \lesssim \|f\|_{L_p^2}$$

3. *$\mathrm{T}\, f \in H_p^{3/2-\epsilon}$ for any $\epsilon > 0$ and*
$$\|\mathrm{T}\, f\|_{H_p^{3/2-\epsilon}} \lesssim \|f\|_{L_p^2}$$

Proof. Let $f \in H_p^s$ for some $s \geq 0$. By the definition of T (see (5.3)) we have $\mathrm{T}\, f \in H_p^1$ (T exists and is well-defined by Lax-Milgram). Therefore, $\mathrm{T}\, f$ is continuous (Theorem 3.27). Let $j \in \{1, \ldots, J\}$. Since $f \in H_p^s$, we have $f_j \in H^s(\Omega_j)$. From (5.3) and since $\mathrm{T}\, f$ is continuous and n_j^2 is constant on Ω_j, we also have that $w_j = u_j$ is a weak solution to the boundary value problem,

$$\begin{aligned} L_j w_j &= h_j \qquad \text{in } \Omega_j \\ w_j &= \mathrm{T}\, f \qquad \text{on } \partial\Omega_j \end{aligned} \tag{5.9}$$

where $L_j := -\frac{1}{n_j^2}(\frac{d}{dx} + i\xi)^2 + (\frac{K}{n_j^2} - c)$ and $h_j := \frac{1}{n_j^2} f_j$. Therefore, with equality in the distributional sense, we have

$$u_j'' = -2i\xi u_j' + (\xi^2 + cn_j^2 - K)u_j + f_j$$

and so, taking the $\|\cdot\|_{H^s(\Omega_j)}$ norm and using the triangle inequality, we get

$$\|u_j\|_{H^{s+2}(\Omega_j)} \lesssim \|u_j\|_{H^{s+1}(\Omega_j)} + \|f_j\|_{H^s(\Omega_j)} \tag{5.10}$$

The result of Part 1 for $s = 0$ then follows from (5.10) using $\|u_j\|_{H^1(\Omega_j)} \lesssim \|f_j\|_{L^2(\Omega_j)}$ (Lax-Milgram). We can then prove Part 1 for $s \in \mathbb{R}$, $s > 0$ by using the following inductive argument.

First, we prove that Part 1 is true for $s \in \mathbb{R}$, $0 \leq s \leq 1$. Equation (5.10) implies

that

$$\begin{aligned}\|u_j\|_{H^{s+2}(\Omega_j)} &\lesssim \|u_j\|_{H^{s+1}(\Omega_j)} + \|f_j\|_{H^s(\Omega_j)} \\ &\leq \|u_j\|_{H^2(\Omega_j)} + \|f_j\|_{H^s(\Omega_j)} \qquad \text{since } s \leq 1 \\ &\lesssim \|f_j\|_{L^2(\Omega_j)} + \|f_j\|_{H^s(\Omega_j)} \qquad \text{by Part 1 with } s = 0 \\ &\lesssim \|f_j\|_{H^s(\Omega_j)}\end{aligned} \tag{5.11}$$

Now assume that Part 1 is true for $s \in \mathbb{R}$, $0 \leq s \leq t$ for some $t \in \mathbb{N}$ (IH). Let $s \in [t, t+1]$. Then, using (5.10), we get

$$\begin{aligned}\|u_j\|_{H^{s+2}(\Omega_j)} &\lesssim \|u_j\|_{H^{s+1}(\Omega_j)} + \|f_j\|_{H^s(\Omega_j)} \\ &\lesssim \|f_j\|_{H^{s-1}(\Omega_j)} + \|f_j\|_{H^s(\Omega_j)} \qquad \text{by (IH)} \\ &\lesssim \|f_j\|_{H^s(\Omega_j)}.\end{aligned} \tag{5.12}$$

Therefore, Part 1 is true for $s \in \mathbb{R}$, $s \geq 0$ by induction using (5.11) and (5.12).

Part 2. Part 1 implies that $u_j \in H^2(\Omega_j)$. Theorem 3.27 then implies that $u_j \in C^1(\Omega_j)$ and $\frac{1}{n_j^2}(\frac{d}{dx} + i\xi)u_j \in C(\Omega_j)$ for each $j = 1, \ldots, J$ since the n_j^2 are constants. Therefore, to show that $\frac{1}{n^2}(\frac{d}{dx} + i\xi)\,\mathrm{T}\,f \in C_p(\Omega)$ we only need to consider $\frac{1}{n^2}(\frac{d}{dx} + i\xi)\,\mathrm{T}\,f(x)$ at $x = x_j$ for $j = 1, \ldots, J$.

Fix $j \in \{1, 2, \ldots, J\}$. We will show that $\frac{1}{n^2}(\frac{d}{dx} + i\xi)\,\mathrm{T}\,f(x)$ is continuous at $x = x_j$ via an arguement similar to that used on page 582 of [12]. But first, we multiply $\mathrm{T}\,f$ by a cut-off function $\psi \in C^\infty(\mathbb{R})$ so that $\operatorname{supp}\psi\,\mathrm{T}\,f \subset\subset (x_{j-1}, x_{j+1})$ and $\frac{1}{n^2}(\frac{d}{dx} + i\xi)(\psi\,\mathrm{T}\,f)$ is continuous for all $x \in \mathbb{R}\backslash\{x_j\}$

We define $\psi \in C^\infty(\mathbb{R})$ in the following way, define the open interval $I_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ and set $\psi := J_\delta * 1_{I_j}$ (recall definition of the usual mollifier function J_δ from Subsection 3.1.5) where $0 < \delta < \frac{1}{2}\min\{|\Omega_{j-1}|, |\Omega_j|\}$ and $1_{I_j}(x)$ is the characteristic function for I_j. By our definition we have $\psi(x_k) = \delta_{jk}$ (Kronecker delta) for $k = 1, \ldots, J$.

Using the product rule, the definition of T (see (5.3)) and the fact that ψ is real-

valued, we can write

$$\begin{aligned}
a(\psi \,\mathrm{T}\, f, \phi) &= \int_\Omega \tfrac{1}{n^2} \left(\left[(\tfrac{d}{dx} + i\xi)(\psi \,\mathrm{T}\, f) \right] \overline{(\tfrac{d}{dx} + i\xi)\phi} + (\mathrm{K} - cn^2)(\psi \,\mathrm{T}\, f)\overline{\phi} \right) dx \\
&= \int_\Omega \tfrac{1}{n^2} \left(\left[(\tfrac{d}{dx} + i\xi) \,\mathrm{T}\, f \right] \overline{(\tfrac{d}{dx} + i\xi)\phi} + (\mathrm{K} - cn^2)(\psi \,\mathrm{T}\, f)\overline{\phi} \right) dx \\
&\quad + \int_\Omega \tfrac{d\psi}{dx} \tfrac{1}{n^2} \,\mathrm{T}\, f \overline{(\tfrac{d}{dx} + i\xi)\phi} dx \\
&= \int_\Omega \tfrac{1}{n^2} \left(\left[(\tfrac{d}{dx} + i\xi) \,\mathrm{T}\, f \right] \overline{(\tfrac{d}{dx} + i\xi)(\psi\phi)} + (\mathrm{K} - cn^2) \,\mathrm{T}\, f \overline{(\psi\phi)} \right) dx \\
&\quad + \int_\Omega \tfrac{d\psi}{dx} \tfrac{1}{n^2} \left(\mathrm{T}\, f \overline{(\tfrac{d}{dx} + i\xi)\phi} - \left[(\tfrac{d}{dx} + i\xi) \,\mathrm{T}\, f \right] \overline{\phi} \right) dx \\
&= b(f, \psi\phi) + \int_\Omega \tfrac{d\psi}{dx} \tfrac{1}{n^2} \left(\mathrm{T}\, f \overline{(\tfrac{d}{dx} + i\xi)\phi} - \left[(\tfrac{d}{dx} + i\xi) \,\mathrm{T}\, f \right] \overline{\phi} \right) dx \\
&= b(\psi f, \phi) + \int_\Omega \tfrac{d\psi}{dx} \tfrac{1}{n^2} \left(\mathrm{T}\, f \overline{(\tfrac{d}{dx} + i\xi)\phi} - \left[(\tfrac{d}{dx} + i\xi) \,\mathrm{T}\, f \right] \overline{\phi} \right) dx \quad \forall \phi \in H^1_p.
\end{aligned} \tag{5.13}$$

For every $k \in \{1, \ldots, J\}$ we find that, by restricting the choice of $\phi \in H^1_p$ so that $\phi \in C^\infty_p$ and $\mathrm{supp}(\phi|_\Omega) \subset\subset \Omega_j$, (5.13) implies that

$$\begin{aligned}
\int_{\Omega_k} \tfrac{1}{n_k^2} (\tfrac{d}{dx} + i\xi)(\psi u_k) \overline{(\tfrac{d}{dx} + i\xi)\phi} + \left(\tfrac{K}{n_k^2} - c \right) (\psi u_k)\overline{\phi} dx &= \int_{\Omega_k} \tfrac{1}{n_k^2} (\psi u_k)\overline{\phi} dx + \\
\int_{\Omega_k} \tfrac{d\psi}{dx} \tfrac{1}{n_k^2} \left(u_k \overline{(\tfrac{d}{dx} + i\xi)\phi} - (\tfrac{d}{dx} + i\xi) u_k \overline{\phi} \right) dx \qquad \forall \phi \in C_0^\infty(\Omega_k).
\end{aligned}$$

From Part 1 and Lemma 3.28 we have $\psi u_k \in H^2(\Omega_j)$. Therefore, we may apply integration by parts to get

$$\begin{aligned}
\int_{\Omega_k} \left(-(\tfrac{d}{dx} + i\xi) \tfrac{1}{n_k^2} (\tfrac{d}{dx} + i\xi)(\psi u_k) + (\tfrac{K}{n_k^2} - c)\psi u_k \right) \overline{\phi} dx &= \int_{\Omega_k} \tfrac{1}{n_k^2} (\psi u_k)\overline{\phi} dx \\
+ \int_{\Omega_k} \left(-\tfrac{1}{n_k^2} (\tfrac{d}{dx} + i\xi)(\tfrac{d\psi}{dx} u_k) - \tfrac{d\psi}{dx} \tfrac{1}{n_k^2} (\tfrac{d}{dx} + i\xi) u_k \right) \overline{\phi} dx \qquad \forall \phi \in C_0^\infty(\Omega_k).
\end{aligned} \tag{5.14}$$

Since (5.14) is true for all $k \in \{1, \ldots, J\}$, we get

$$\begin{aligned}
-(\tfrac{d}{dx} + i\xi) \tfrac{1}{n^2} (\tfrac{d}{dx} + i\xi)(\psi \,\mathrm{T}\, f) + (\tfrac{K}{n^2} - c)\psi \,\mathrm{T}\, f &= \tfrac{1}{n^2} (\psi \,\mathrm{T}\, f) \\
+ \left(-\tfrac{1}{n^2} (\tfrac{d}{dx} + i\xi)(\tfrac{d\psi}{dx} \,\mathrm{T}\, f) - \tfrac{d\psi}{dx} \tfrac{1}{n^2} (\tfrac{d}{dx} + i\xi) \,\mathrm{T}\, f \right)
\end{aligned} \tag{5.15}$$

almost everywhere in Ω.

Now let $\phi \in C_0^\infty(\Omega)$ (then $\phi = 0$ on $\partial\Omega$ and it can be extended periodically so that

it is in $C_p^\infty \subset H_p^1$). Using (5.13) and the fact that $\operatorname{supp}\psi \subset\subset [x_{j-1}, x_{j+1}]$ we get

$$\begin{aligned}
b(\psi f,\phi) &+ \int_\Omega \left(-\tfrac{1}{n^2}(\tfrac{d}{dx}+i\xi)(\tfrac{d\psi}{dx}\,\mathrm{T}\,f) - \tfrac{d\psi}{dx}\tfrac{1}{n^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f\right)\overline{\phi}dx \\
&= b(\psi f,\phi) + \int_\Omega \tfrac{d\psi}{dx}\tfrac{1}{n^2}\left(\mathrm{T}\,f\overline{(\tfrac{d}{dx}+i\xi)\phi} - (\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f\overline{\phi}\right)dx \\
&= a(\psi\,\mathrm{T}\,f,\phi) \qquad\qquad \text{by (5.13)} \\
&= \int_{\Omega_{j-1}} \tfrac{1}{n_{j-1}^2}(\tfrac{d}{dx}+i\xi)(\psi u_{j-1})\overline{(\tfrac{d}{dx}+i\xi)\phi} + \left(\tfrac{K}{n_{j-1}^2}-c\right)\psi u_{j-1}\overline{\phi}dx \\
&\quad + \int_{\Omega_j} \tfrac{1}{n_j^2}(\tfrac{d}{dx}+i\xi)(\psi u_j)\overline{(\tfrac{d}{dx}+i\xi)\phi} + \left(\tfrac{K}{n_j^2}-c\right)\psi u_j\overline{\phi}dx \\
&= \left[\tfrac{1}{n_{j-1}^2}(\tfrac{d}{dx}+i\xi)(\psi u_{j-1})\overline{\phi}\right]_{x_{j-1}}^{x_j} \\
&\quad - \int_{\Omega_{j-1}} \left((\tfrac{d}{dx}+i\xi)\tfrac{1}{n_{j-1}^2}(\tfrac{d}{dx}+i\xi)(\psi u_{j-1}) + (\tfrac{K}{n_{j-1}^2}-c)\psi u_{j-1}\right)\overline{\phi}dx \\
&\quad + \left[\tfrac{1}{n_j^2}(\tfrac{d}{dx}+i\xi)(\psi u_j)\overline{\phi}\right]_{x_j}^{x_{j+1}} \\
&\quad - \int_{\Omega_j} \left((\tfrac{d}{dx}+i\xi)\tfrac{1}{n_j^2}(\tfrac{d}{dx}+i\xi)(\psi u_j) + (\tfrac{K}{n_j^2}-c)\psi u_j\right)\overline{\phi}dx \\
&= \lim_{\epsilon_1 \searrow 0}\left(\tfrac{1}{n_{j-1}^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f(x_j-\epsilon_1)\right) - \lim_{\epsilon_1\searrow 0}\left(\tfrac{1}{n_j^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f(x_j+\epsilon_1)\right) \\
&\quad - \int_\Omega \left((\tfrac{d}{dx}+i\xi)\tfrac{1}{n^2}(\tfrac{d}{dx}+i\xi)(\psi\,\mathrm{T}\,f) + (\tfrac{K}{n^2}-c)(\psi\,\mathrm{T}\,f)\right)\overline{\phi}dx.
\end{aligned}$$

By (5.15) and the properties of ψ, this implies that

$$\lim_{\epsilon_1\searrow 0}\tfrac{1}{n_j^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f(x_j+\epsilon_1) = \lim_{\epsilon_1\searrow 0}\tfrac{1}{n_{j-1}^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f(x_j-\epsilon_1).$$

Therefore, $\frac{1}{n^2}(\frac{d}{dx}+i\xi)\,\mathrm{T}\,f(x)$ is continuous at $x=x_j$ and we have now shown that $\frac{1}{n^2}(\frac{d}{dx}+i\xi)\,\mathrm{T}\,f \in C_p(\Omega)$.

We now show that $\|\frac{1}{n^2}(\frac{d}{dx}+i\xi)\,\mathrm{T}\,f\|_{H_p^1} \lesssim \|f\|_{H_p^2}$. In a distributional sense, the definition of T (see (5.3)) implies

$$-(\tfrac{d}{dx}+i\xi)\tfrac{1}{n^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f + (\tfrac{K}{n^2}-c)\,\mathrm{T}\,f = \tfrac{1}{n^2}f$$

which further implies that

$$-\tfrac{d}{dx}(\tfrac{1}{n^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f) = i\xi\tfrac{1}{n^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f - (\tfrac{K}{n^2}-c)\,\mathrm{T}\,f + \tfrac{1}{n^2}f. \tag{5.16}$$

Therefore, by taking the $\|\cdot\|_{L^2_p}$ of (5.16) and using the triangle inequality we get

$$\begin{aligned}\|\tfrac{1}{n^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f\|_{H^1_p} &\lesssim \|\tfrac{d}{dx}(\tfrac{1}{n^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f)\|_{L^2_p}+\|\tfrac{1}{n^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f\|_{L^2_p}\\ &\lesssim \|\mathrm{T}\,f\|_{H^1_p}+\|\mathrm{T}\,f\|_{L^2_p}+\|f\|_{L^2_p} \qquad \text{by (5.16)}\\ &\lesssim \|f\|_{L^2_p} \qquad \text{by Lax-Milgram.}\end{aligned}$$

The remainder of the result follows from Theorem 3.27.

Part 3. Our proof of $\mathrm{T}\,f \in H^{3/2-\epsilon}_p$ for $\epsilon > 0$ in Part 3 is similar to a proof in [65] and relies on a result in [32]. Instead of showing that $\mathrm{T}\,f \in H^s_p$ for $s < 3/2$, it is sufficient to show that $(\mathrm{T}\,f)' \in H^s_p$ for $s < 1/2$. From Part 1 we have $u_j \in H^2(\Omega_j)$ for every $j = 1,\ldots,J$. This implies that $u_j' \in H^1(\Omega_j) \subset H^s(\Omega_j)$ for $s < 1/2$. Now extend each u_j with zero to all of $\mathbb{R}$. Denote this extension of u_j with $\widetilde{u}_j$. Define $\widetilde{u} = \sum_{j=1}^J \widetilde{u}_j$. A remark after Theorem 1.2.16 in [32] (using Definition 1.2.4 in [32]) says that $u_j' \in H^s(\Omega_j) \Longrightarrow \widetilde{u}_j' \in H^s(\mathbb{R})$ for $0 \le s < 1/2$. By the definition of $\widetilde{u}$ it then follows that $\widetilde{u}' \in H^s(\mathbb{R})$ for $0 \le s < 1/2$. Then, by the definition of $H^s(\Omega)$, we get $\tilde{u}'|_\Omega \in H^s(\Omega)$. But $\mathrm{T}\,f = \tilde{u}|_\Omega$ almost everywhere. Therefore, $(\mathrm{T}\,f)' = \tilde{u}'|_\Omega \in H^s(\Omega)$ for $0 \le s < 1/2$. Theorem 3.29 then implies that $(\mathrm{T}\,f)' \in H^s_p$ for $0 \le s < 1/2$.

To prove the estimate for $\|\mathrm{T}\,f\|_{H^{3/2-\epsilon}_p}$ for $\epsilon > 0$ we use the estimate from Part 2 and the following argument,

$$\begin{aligned}\|\mathrm{T}\,f\|_{H^{3/2-\epsilon}_p} &\lesssim \|(\mathrm{T}\,f)'\|_{H^{1/2-\epsilon}_p}+|[\mathrm{T}\,f]_0| && \text{by definition of } \|\cdot\|_{H^s_p}\\ &= \|(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f - i\xi\,\mathrm{T}\,f\|_{H^{1/2-\epsilon}_p}+|[\mathrm{T}\,f]_0| &&\\ &\lesssim \|(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f\|_{H^{1/2-\epsilon}_p}+\|\mathrm{T}\,f\|_{H^{1/2-\epsilon}_p} && \text{by triangle inequality}\\ &\lesssim \|n^2\|_{H^{1/2-\epsilon}_p}\|\tfrac{1}{n^2}(\tfrac{d}{dx}+i\xi)\,\mathrm{T}\,f\|_{H^1_p}+\|\mathrm{T}\,f\|_{H^1_p} && \text{by Theorem 3.28}\\ &\lesssim \|n^2\|_{H^{1/2-\epsilon}_p}\|f\|_{L^2_p}+\|\mathrm{T}\,f\|_{H^1_p} && \text{by Part 2}\\ &\lesssim \|n^2\|_{H^{1/2-\epsilon}_p}\|f\|_{L^2_p}+\|f\|_{L^2_p} && \text{by Lax-Milgram}\\ &\lesssim \|f\|_{L^2_p} && \text{by Theorem 3.40.}\end{aligned}$$

□

We now present a corollary to Theorem 5.4 for eigenfunctions of Problem 5.1. The proof is an elementary application of Theorem 5.4.

Corollary 5.5. *Let u be an eigenfunction of Problem 5.1 and define $u_j := u|_{\Omega_j}$ for each $j = 1,\ldots,J$. Then*

1. *$u_j \in C^\infty(\Omega_j)$ for each $j = 1,\ldots,J$.*

2. $\frac{1}{n^2}(\frac{d}{dx}+i\xi)u \in H_p^1$ *(and is continuous by Theorem 3.27) and*

$$\|\tfrac{1}{n^2}\left(\tfrac{d}{dx}+i\xi\right)u\|_\infty \lesssim \|\tfrac{1}{n^2}\left(\tfrac{d}{dx}+i\xi\right)u\|_{H_p^1} \lesssim \|u\|_{H_p^1}$$

3. $u \in H_p^{3/2-\epsilon}$ *for any* $\epsilon > 0$ *and*

$$\|u\|_{H_p^{3/2-\epsilon}} \lesssim \|u\|_{H_p^1}$$

Using these regularity results we can derive the following approximation error results for plane waves. Recall the definition of $\mathcal{S}_G \subset H_p^1$ for $G \in \mathbb{N}$,

$$\mathcal{S}_G := \operatorname{span}\{\mathrm{e}^{i2\pi gx} : g \in \mathbb{Z}, |g| \le G\}.$$

Corollary 5.6. *Using Theorem 5.4 we get the following two corollary results:*

1. *If* $u \in H_p^1$ *then*

$$\inf_{\chi\in\mathcal{S}_G} \|\operatorname{T} u - \chi\|_{H_p^1} \lesssim G^{-1/2+\epsilon}\|u\|_{H_p^1} \qquad \forall\epsilon > 0.$$

2. *If* u *is an eigenfunction of Problem 5.1 then*

$$\inf_{\chi\in\mathcal{S}_G} \|u - \chi\|_{H_p^1} \lesssim G^{-1/2+\epsilon}\|u\|_{H_p^1} \qquad \forall\epsilon > 0.$$

Proof. Part 1. Let $u \in H_p^1$ and $\epsilon > 0$. Then, by choosing $\chi = \operatorname{P}_G^{(\mathcal{S})}\operatorname{T} u$ (where $\operatorname{P}_G^{(\mathcal{S})}$ is defined in Subsection 3.2.5) we get

$$\begin{aligned}
\inf_{\chi\in\mathcal{S}_G} \|\operatorname{T} u - \chi\|_{H_p^1} &\le \|\operatorname{T} u - \operatorname{P}_G^{(\mathcal{S})}\operatorname{T} u\|_{H_p^1} \\
&\le G^{-1/2+\epsilon}\|\operatorname{T} u\|_{H_p^{3/2-\epsilon}} && \text{by Lemma 3.30} \\
&\lesssim G^{-1/2+\epsilon}\|u\|_{H_p^1} && \text{by Part 3 of Theorem 5.4.}
\end{aligned}$$

Part 2 follows directly from Part 1. □

5.3.2 Spectral Galerkin Method

Before considering the errors for the plane wave expansion method let us first consider the spectral Galerkin method applied to Problem 5.1. As we discussed at the beginning of Section 5.2 this method is *not* the plane wave expansion method (we will prove this in the next subsection) and it does not share the computational efficiencies of the plane wave expansion method (unlike for the 1D TE Mode Problem where the these two methods are the same). It does, however, allow us to apply all of the error

analysis techniques from [6] that we used in Subsection 4.2.3 to develop a complete error analysis.

Applying the spectral Galerkin method with finite dimensional subspace $\mathcal{S}_G$ to Problem 5.1 yields the following discrete variational eigenvalue problem.

Problem 5.7. Find $\lambda_G \in \mathbb{R}$ and $0 \neq u_G \in \mathcal{S}_G$ such that

$$a(u_G, v_G) = \lambda_G b(u_G, v_G) \qquad \forall v_G \in \mathcal{S}_G.$$

This finite dimensional problem is equivalent to a matrix eigenproblem and matrix-vector products can be computed in $\mathcal{O}(N \log N)$ operations using the Fast Fourier Transform, but the 2nd-order part of the differential operator does not reduce to a simple diagonal matrix and we do not have an optimal preconditioner for solving linear systems.

The first step of the error analysis is to define the solution operator $\mathrm{T}_G : L^2_p \to \mathcal{S}_G$ that is associated with Problem 5.7. For $f \in L^2_p$ we define $\mathrm{T}_G f$ by

$$a(\mathrm{T}_G f, v_G) = b(f, v_G) \qquad \forall v_G \in \mathcal{S}_G.$$

Note that the definition of T_G is similar to the definition of T_n in (3.43). Recall that T is the solution operator associated with Problem 5.1 (see (5.3)). The following lemma proves some properties of T_G.

Lemma 5.8. *The following properties hold for* T *and* T_G.

1. $\mathrm{T}_G : H^1_p \to H^1_p$ *is bounded, compact and self-adjoint with respect to* $a(\cdot, \cdot)$.
2. *For* $\epsilon > 0$,
$$\| \mathrm{T} - \mathrm{T}_G \|_{H^1_p} \lesssim G^{-1/2+\epsilon}.$$

Proof. The proof of Part 1 is the same as the proof of Part 2 of Lemma 4.23, whereas the proof of Part 2 follows from Corollary 5.6 using Part 2 of Lemma 3.74. □

Now we use Theorem 3.68 to prove the main result of this subsection.

Theorem 5.9. *Let* λ *be an eigenvalue of Problem 5.1 with multiplicity* m *and corresponding eigenspace* M. *Then, for sufficiently large* G *and arbitrarily small* $\epsilon > 0$ *there exist* m *eigenvalues* $\lambda_1(G), \lambda_2(G), \ldots, \lambda_m(G)$ *of Problem 5.7 (counted according to their multiplicity) with corresponding eigenspaces* $M_1(\lambda_1), \ldots, M_m(\lambda_m)$ *and a space*

$$\mathcal{M}_G = \bigoplus_{j=1}^{m} M_j(\lambda_j)$$

such that

$$\delta(M, \mathcal{M}_G) \lesssim G^{-1/2+\epsilon}$$

and

$$|\lambda - \lambda_j| \lesssim G^{-1+\epsilon} \qquad \textit{for } j = 1, \ldots, m.$$

Proof. For the proof of this result we would like to apply Theorem 3.68. We have already defined the solution operator T that is associated with Problem 5.1. From Theorem 5.3 we know that T is bounded, compact, and self-adjoint with respect to $a(\cdot, \cdot)$. From Lemma 5.8 we know that T_G for $G \in \mathbb{N}$ are a family of bounded, compact, self-adjoint operators such that $\| \mathrm{T} - \mathrm{T}_G \|_{H^1_p} \to 0$ as $G \to \infty$. The result then follows by applying Theorem 3.68 and Lemma 3.74. □

So, we see that the error analysis for Problem 5.7 is the same as for the Scalar 2D Problem and the 1D TE Mode Problem. We have shown that the eigenfunction error is optimal in the sense that it decays at the same rate as the approximation error of $\mathcal{S}_G$ approximating exact eigenfunctions and the approximation error decay rate depends on the regularity of the exact eigenfunctions. Therefore, the limiting factor for the spectral Galerkin method applied to the 1D TM Mode Problem is the regularity of the exact eigenfunctions, and because the eigenfunctions of the 1D TM Mode Problem have less regularity than the eigenfunctions of the 1D TE Mode Problem, the spectral Galerkin method converges at a slower rate for the 1D TM Mode Problem than for the 1D TE Mode Problem. We have also shown that the eigenvalues converge at twice the rate of the eigenfunctions as we did for the spectral Galerkin method applied to the 1D TE Mode Problem. This property is the same for the TE and TM Mode Problems because they are both self-adjoint and they both possess "Galerkin orthogonality".

Now we will consider the plane wave expansion method. One of the first things we prove is that the plane wave expansion method is not equivalent to the spectral Galerkin method for the 1D TM Mode Problem.

5.3.3 Plane Wave Expansion Method

In this subsection we attempt to analyse the errors of the plane wave expansion method applied to the 1D TM Mode Problem. The presentation of the plane wave expansion method that we gave in Subsection 5.2 is the same as that used in [64] and [39] and does not lend itself easily to our error analysis approach. For the error analysis we attempt to write down a discrete variational eigenproblem that is equivalent to (5.8). In this subsection we begin by defining two discrete variational problems that are equivalent to (5.8).

Unfortunately, neither of these discrete variational eigenproblems are equivalent to the spectral Galerkin method (Problem 5.7) and we can not use the error analysis from the previous subsection for the plane wave expansion method. Attempting to analyse the error using other theoretical techniques has also failed so far for both of our discrete variational eigenproblems, as we explain.

Without a complete error analysis for the plane wave expansion method we will use the approximation error result that we proved in Corollary 5.6 for eigenfunctions of Problem 5.1 approximated by plane waves. This estimate gives us an upper limit for the rate at which the plane wave expansion method can converge for the eigenfunctions of Problem 5.1. In the next subsection we will see that for our numerical examples, the plane wave expansion method actually achieves this fastest possible convergence rate for the eigenfunctions and we conclude that we should be able to prove that the planeave expansion method is stable and that, as in all other cases, the limiting factor for the method is the regularity of the eigenfunctions of Problem 5.1.

We will need to define the following two finite dimensional function spaces. For the same $G \in \mathbb{N}$, define

$$\mathcal{S}_G := \mathcal{S}_G^{(1)} = \operatorname{span}\{\mathrm{e}^{i2\pi gx} : |g| \leq G\}$$
$$\mathcal{S}_{G\star} := \operatorname{span}\{n^2(x)\,\mathrm{e}^{i2\pi gx} : |g| \leq G\}.$$

We have $N = \dim \mathcal{S}_G = \dim \mathcal{S}_{G\star} = 2G + 1$. Note that we have already used $\mathcal{S}_G$ many times throughout this thesis but we have not seen $\mathcal{S}_{G\star}$ before.

Now we define two discrete variational eigenproblems and prove that they are both equivalent to (5.8) (see Lemma 5.12 below).

Problem 5.10. Find $\lambda_G \in \mathbb{R}$ and $0 \neq u_G \in \mathcal{S}_G$ such that

$$a_1(u_G, v_G) = \lambda_G b_1(u_G, v_G) \qquad \forall v_G \in \mathcal{S}_G$$

where

$$a_1(u_G, v_G) = \int_\Omega \left(\tfrac{d}{dx} + i\xi\right) u_G \overline{\left(\tfrac{d}{dx} + i\xi\right) v_G} + (\log n^2)' \left(\tfrac{d}{dx} + i\xi\right) u_G \overline{v_G} + (K - \gamma) u_G \overline{v_G} dx$$
$$b_1(u_G, v_G) = \int_\Omega u_G \overline{v_G} dx.$$

Problem 5.11. Find $\lambda_G \in \mathbb{R}$ and $0 \neq u_G \in \mathcal{S}_G$ such that

$$a(u_G, v_G) = \lambda_G b(u_G, v_G) \qquad \forall v_G \in \mathcal{S}_{G\star}.$$

In Problem 5.10 it is not entirely clear how $a_1(\cdot,\cdot)$ is defined because $(\log n^2)'$ is not a classical function. It is a derivative of a discontinuous function and we interpret it in the following way. For any $f \in \mathcal{D}_p'(\mathbb{R})$ (i.e. f is a periodic distribution), Theorem 3.22 ensures that f has a Fourier Series and we get

$$\int_\Omega f\phi dx = \int_\Omega (\mathrm{P}_G^{(\mathcal{S})} f)\phi dx \qquad \forall \phi \in \mathcal{S}_G$$

where the projection $\mathrm{P}_G^{(\mathcal{S})}$ is defined in Subsection 3.2.5. Therefore,

$$\int_\Omega (\log n^2)'(\tfrac{d}{dx}+i\xi)u_G\overline{v_G}dx = \int_\Omega (\mathrm{P}_{2G}^{(\mathcal{S})}(\log n^2)')(\tfrac{d}{dx}+i\xi)u_G\overline{v_G}dx \qquad \forall u_G, v_G \in \mathcal{S}_G.$$

Now we show that Problems 5.10 and 5.11 are both representations of the plane wave expansion method applied to the 1D TM Mode Problem by showing that they are equivalent to the matrix eigenproblem (5.8).

Lemma 5.12. *Problem 5.10, Problem 5.11 and* (5.8) *are equivalent problems.*

Proof. First, we show that Problem 5.10 is equivalent to Problem 5.11. We need to recognise that $V_G = \{\mathrm{e}^{i2\pi gx} : g \in \mathbb{Z}, |g| \le G\}$ is a basis for $\mathcal{S}_G$ and $V_{G\star} = \{n^2(x)\,\mathrm{e}^{i2\pi gx} : g \in \mathbb{Z}, |g| \le G\}$ is a basis for $\mathcal{S}_{G\star}$. Then, (λ_G, u_G) is an eigenpair of Problem 5.10 if and only if

$$\begin{aligned}
& a_1(u_G, v_G) = \lambda_G b_1(u_G, v_G) && \forall v_G \in V_G \\
\Leftrightarrow \quad & \int_\Omega (\tfrac{d}{dx}+i\xi)u_G\overline{(\tfrac{d}{dx}+i\xi)v_G} + \tfrac{(n^2)'}{n^2}(\tfrac{d}{dx}+i\xi)u_G\overline{v_G} \\
& \qquad +(\mathrm{K}-\gamma)u_G\overline{v_G}dx = \lambda_G\int_\Omega u_G\overline{v_G}dx && \forall v_G \in V_G \\
\Leftrightarrow \quad & \int_\Omega \tfrac{1}{n^2}\big((\tfrac{d}{dx}+i\xi)u_G\overline{(n^2(\tfrac{d}{dx}+i\xi)v_G + (n^2)'v_G)} \\
& \qquad +(\mathrm{K}-\gamma)u_G\overline{(n^2v_G)}\big)dx = \lambda_G\int_\Omega \tfrac{1}{n^2}u_G\overline{(n^2v_G)}dx && \forall v_G \in V_G \\
\Leftrightarrow \quad & \int_\Omega \tfrac{1}{n^2}\big((\tfrac{d}{dx}+i\xi)u_G\overline{(\tfrac{d}{dx}+i\xi)(n^2v_G)} \\
& \qquad +(\mathrm{K}-\gamma)u_G\overline{(n^2v_G)}\big)dx = \lambda_G\int_\Omega \tfrac{1}{n^2}u_G\overline{(n^2v_G)}dx && \forall v_G \in V_G \\
\Leftrightarrow \quad & \int_\Omega \tfrac{1}{n^2}\big((\tfrac{d}{dx}+i\xi)u_G\overline{(\tfrac{d}{dx}+i\xi)w_G} \\
& \qquad +(\mathrm{K}-\gamma)u_G\overline{w_G}\big)dx = \lambda_G\int_\Omega \tfrac{1}{n^2}u_G\overline{w_G}dx && \forall w_G \in V_{G\star} \\
& a(u_G, w_G) = \lambda_G b(u_G, w_G) && \forall w_G \in V_{G\star}
\end{aligned}$$

if and only if (λ_G, u_G) is an eigenpair of Problem 5.11. Therefore, Problem 5.10 is equivalent to Problem 5.11.

To complete the proof we will now show that Problem 5.10 is equivalent to (5.8). Note first that the entries of A in (5.8) satisfy

$$\mathrm{A}_{jk} := a_1(\mathrm{e}^{i2\pi g'x}, \mathrm{e}^{i2\pi gx}) \qquad g, g' = -G, \ldots, G.$$

Now suppose that (λ_G, u_G) is an eigenpair of Problem 5.10. Expand u_G as

$$u_G(x) = \sum_{|h|\leq G} [u_G]_h \, \mathrm{e}^{i2\pi hx}$$

and define a vector $\mathbf{u}$ with entries $u_h = [u_G]_h$ for $h = -G, \ldots, G$. Then (λ_G, u_G) is an eigenpair of Problem 5.10 if and only if

$$\begin{aligned}
& a_1(u_G, \mathrm{e}^{i2\pi gx}) = \lambda_G b_1(u_G, \mathrm{e}^{i2\pi gx}) && \forall g = -G, \ldots, G \\
\Leftrightarrow \quad & \sum_{|h|\leq G} a_1(\mathrm{e}^{i2\pi hx}, \mathrm{e}^{i2\pi gx})[u_G]_h = \lambda_G \sum_{|h|\leq G} b_1(\mathrm{e}^{i2\pi hx}, \mathrm{e}^{i2\pi gx})[u_G]_h && \forall g = -G, \ldots, G \\
\Leftrightarrow \quad & \sum_{|h|\leq G} a_1(\mathrm{e}^{i2\pi hx}, \mathrm{e}^{i2\pi gx})[u_G]_h = \lambda_G [u_G]_g && \forall g = -G, \ldots, G \\
\Leftrightarrow \quad & \sum_{|h|\leq G} \mathrm{A}_{gh} \, u_h = \lambda_G u_g && \forall g = -G, \ldots, G
\end{aligned}$$

if and only if $(\lambda, \mathbf{u})$ is an eigenpair of (5.8). □

Now we consider the error analysis for Problems 5.10 and 5.11 as approximations to Problem 5.1. First, we consider the error analysis for Problem 5.10. The difficulty with using Problem 5.10 is two-fold. The first problem is that $a_1(\cdot, \cdot)$ is not defined on $H_p^1 \times H_p^1$. This is because

$$\int_\Omega (\log n^2)'(\tfrac{d}{dx} + i\xi) u \overline{v} dx$$

is not defined for all $u, v \in H_p^1$. However, as noted after the definition of Problems 5.10 and 5.11 we can replace $(\log n^2)'$ with $\mathrm{P}_{2G}^{(\mathcal{S})}(\log n^2)'$ in $a_1(\cdot, \cdot)$. Unfortunately, this leads to the second difficulty. The new $a_1(\cdot, \cdot)$ (with $\mathrm{P}_{2G}^{(\mathcal{S})}(\log n^2)'$ instead of $(\log n^2)'$) is not bounded independently of G on H_p^1 and we can not prove that it is coercive on H_p^1. Consequently, when we try to apply our usual theory we find that we can not prove that the error will decrease as we increase G.

Now consider Problem 5.11. Since $\mathcal{S}_{G\star} \not\subseteq H_p^1$, Problem 5.11 corresponds to a non-conforming Petrov-Galerkin method applied to Problem 5.1. Although we have not been successful with developing the error analysis in this case, we think that representing the plane wave expansion method in this way, as a non-conforming Petrov-Galerkin method, might be amenable to theory such as that in [85], but this requires further investigation.

In the absence of a complete error analysis for the plane wave expansion method we assume that the method is stable and use the approximation error result from Corollary 5.6 to predict the rate at which the plane wave expansion method should converge. Using Corollary 5.6 we predict that the H_p^1 norm of the eigenfunction error should decay with $\mathcal{O}(G^{-1/2+\epsilon})$ for arbitrarily small $\epsilon > 0$. The numerical results in

Section 5.4 suggest that our assumption that the method is stable is justified and we actually achieve a convergence rate of $\mathcal{O}(G^{-1/2+\epsilon})$ for arbitrarily small $\epsilon > 0$ for the eigenfunctions.

5.4 Examples

In this section we compute approximations to the 1D TM Mode Problem using the plane wave expansion method. We will be solving (5.8) as an approximation to Problem 5.1. We observe that the eigenfunction error decays at the same rate as the approximation error estimate that we proved in Corollary 5.6. This confirms that the plane wave expansion method is stable for these examples and the convergence rate is entirely dependent on the regularity of the eigenfunctions of Problem 5.1. We also observe that the eigenvalues decay at twice the rate of the eigenfunctions. This agrees with the analysis of the spectral Galerkin method that we proved in Subsection 5.3.2. Even though (5.8) is a non-symmentric eigenvalue problem there still appears to be sufficient symmetry in the plane wave expansion method so that the eigenvalues to converge at twice the rate of the eigenfunctions.

We do computations for the PCF structures of Model Problems 1 and 2 that we defined in Subsection 4.1.7 for the 1D TE Mode Problem. In particular, $n(x)$ is a piecewise constant function where $n(x) = 1$ in the *air* regions and $n(x) = 1.4$ in the *glass* regions. Figure 4-1 represents the structure of $n(x)$. As in Chapter 4, $\lambda_0 = \frac{1}{2}$ and there is a 50:50 glass to air ratio. In Figure 5-1 we have plotted the band structure of the spectrum for Model Problems 1 and 2. We see that the band structure is very similar to that of the 1D TE Mode Problem, see Figure 4-3. In Figure 5-1, each band is constructed by projecting the corresponding line onto the vertical axis. And each line is an eigenvalue of (5.8) as a function of $\xi \in B$, i.e. $\lambda(\xi)$. Problem 1 has five bands in $[0, \infty)$. Problem 2 has approximately the same band gaps as Problem 1 and there do not appear to be any obviously isolated eigenvalues. For each band in Problem 1 there are approximately 13 bands in Problem 2. This number corresponds to the number of cells in the supercell of Problem 2. There are small band gaps between every band of Problem 2 but these small gaps arise from having a supercell with finite cladding.

To examine the convergence of the plane wave expansion method we solve (5.8) over a range of values of G. We calculate the error by comparing our eigenvalues and eigenvectors against a reference solution, which is computed by solving (5.8) with $G = 2^{18} - 1$. In Figures 5-2 and 5-3 we see that the errors of the normalised eigenfunctions measured in $\|\cdot\|_{H^1_p}$ decay with $\mathcal{O}(G^{-1/2})$. This is the fastest rate of decay that we could have expected given the approximation error result that we proved in Corollary 5.6. We recall that this approximation error result was limited by the regularity of the exact eigenfunctions. Thus, the rate at which the eigenfunction error decays appears

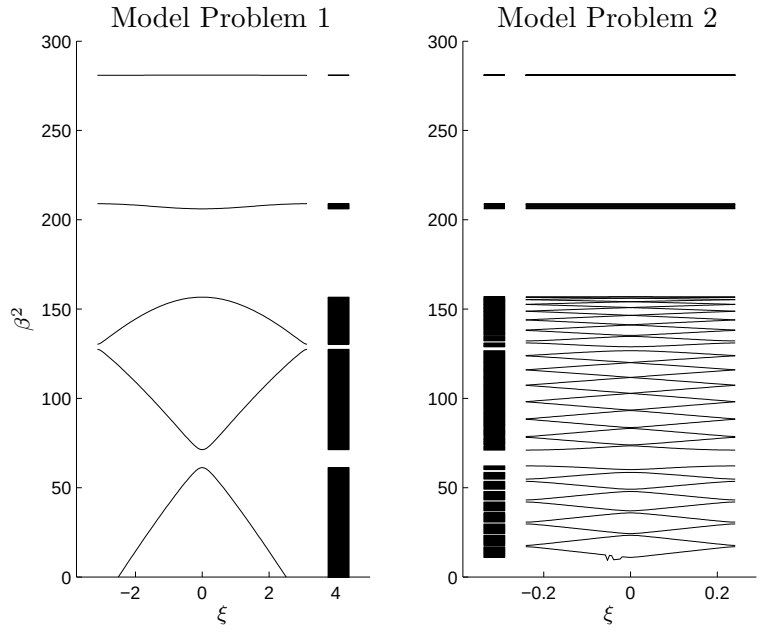

Figure 5-1: A plot of the spectra of Model Problems 1 and 2. The spectra are represented with solid black blocks (or bands) running vertically nearest the middle of the page.

to entirely depend on the regularity of the exact problem. The numerically observed rate of $\mathcal{O}(G^{-1/2})$ for the eigenfunction error is also the same as the convergence rate that we were able to prove for the spectral Galerkin method in Subsection 5.3.2.

In Figures 5-2 and 5-3 we also observe that the relative errors of the eigenvalues are $\mathcal{O}(G^{-1})$. This rate of decay is twice as fast as the decay rate for the eigenfunctions. We managed to prove a similar result for the spectral Galerkin method applied to Problem 5.1 in Subsection 5.3.2, and the proof depended on the self-adjointness of Problem 5.1 as well as on the self-adjointness of Problem 5.7. We also proved and observed this phenomenon in Chapter 4 for the plane wave expansion method applied to the 1D TE Mode Problem and the Scalar 2D Problem, where the proof also depended on the self-adjointness of the continuous and discrete problems. The fact that it also seems to be the case for the plane wave expansion method applied to Problem 5.1 suggests that it might be possible to reformulate 5.8 as a symmetric eigenvalue problem.

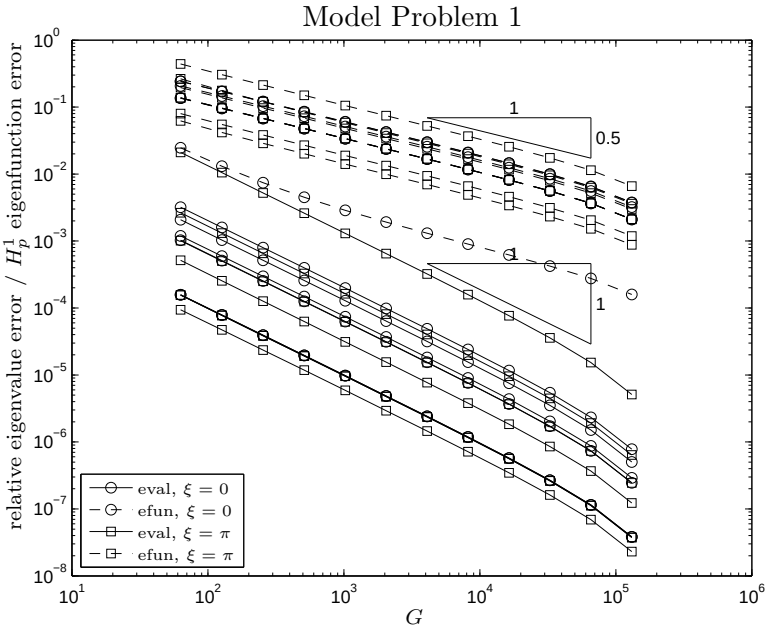

Figure 5-2: Plot of the relative eigenvalue error (eval) and the eigenfunction error measured in the H_p^1 norm (efun) vs. G for the first 5 eigenpairs of Model Problem 1 (solved for both $\xi = 0$ and $\xi = \pi$).

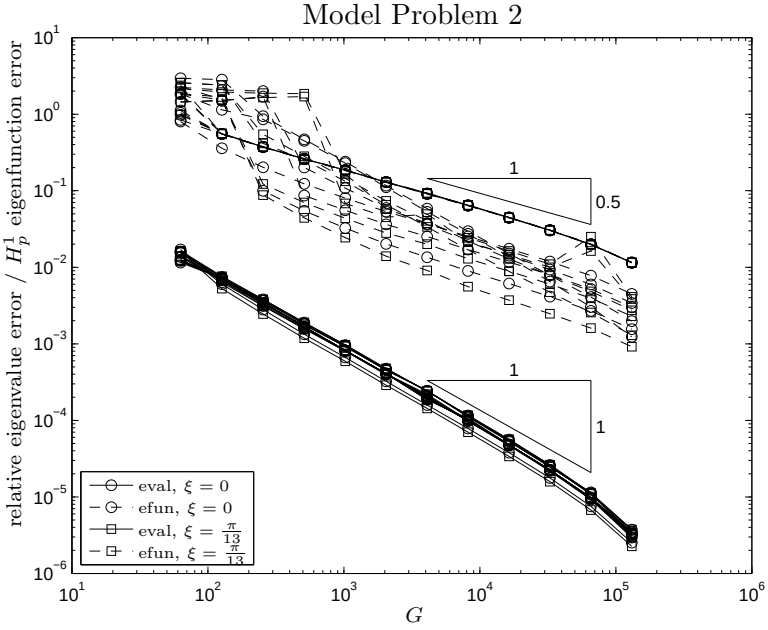

Figure 5-3: Plot of the relative eigenvalue error (eval) and the eigenfunction error measured in the H_p^1 norm (efun) vs. G for the 21st-30th eigenpairs of Model Problem 2 (solved for both $\xi = 0$ and $\xi = \frac{\pi}{13}$).

5.5 Other Examples: Smoothing and Sampling

Although we have not mentioned it yet for the 1D TM Mode Problem we can apply smoothing and/or sampling within the plane wave expansion method, as in Sections 4.3 and 4.4, by modifying the Fourier coefficients of $n^2(x)$ and $(\log n^2)'$. We are interested to see whether or not our conclusions about smoothing and sampling from Chapter 4 for the smoothing and sampling methods applied to the 1D TE Mode Problem and the Scalar 2D Problem are also true for the 1D TM Mode Problem. In particular, we would like to know if smoothing will help the plane wave expansion method and what grid-spacing we should choose in our sampling grid to recover the accuracy of exact Fourier coefficients.

First, we consider the smoothing method. To apply this method we solve (5.8) with $[\gamma]_j$ and $[\log n^2]_j$ in the definition of A in (5.8) replaced with $\mathrm{e}^{-2\pi^2|j|^2\Delta^2}[\gamma]_j$ and $\mathrm{e}^{-2\pi^2|j|^2\Delta^2}[\log n^2]_j$ respectively, where Δ is the parameter that determines the amount of smoothing. In Figure 5-4 we have plotted the errors of the eigenvalues and eigenfunctions for the plane wave expansion method with smoothing with G fixed ($G = 2^{17} - 1$) and varying amounts of smoothing (varying Δ). In this case the reference solution is the solution to (5.8) with $G = 2^{18} - 1$ and $\Delta = 0$ (no smoothing). We see that the error depends on Δ in a more complicated way than for the Scalar 2D Problem and the 1D TE Mode Problem in Section 4.3 (c.f. Figure 4-15). There appear to be two "regimes" for how the error depends on Δ. Here, we will discuss the eigenfunction errors because the error dependence on Δ is clearer in this case than in the case of the eigenvalue errors. For $\Delta \in [10^{-7}, 10^{-5}]$ the eigenfunction errors appear to have $\mathcal{O}(\Delta^{3/2})$ dependence on Δ. This is the same dependence that we saw for the 1D TE Mode Problem, but for $\Delta > 10^{-3}$ we see that the eigenfunction errors appear to have $\mathcal{O}(\Delta^{1/2})$ dependence on Δ. Although we do not have any rigorous mathematical explanation for this behaviour, one possible explanation is that in the smoothing method we modify A from (5.8) by changing the entries of both W and V, and the changes to W and V are contributing to the error in different ways, resulting in two "regimes". Also, in one of the "regimes" we see the same error behaviour as for the 1D TE Mode Problem. This might be because the matrix V is the same matrix V as was used in the 1D TE Mode Problem.

In Figures 5-5 and 5-6 we have plotted the errors of the plane wave expansion method with smoothing for varying G where we have chosen $\Delta = G^r$ for different constants r. Again, the reference solution is the solution to (5.8) with $G = 2^{18} - 1$ and $\Delta = 0$, i.e. the plane wave expansion method without smoothing. From these plots we conclude that we should choose $\Delta \leq G^{-3/2}$ to recover the convergence rate that we see for the plane wave expansion method *without* smoothing and as before, smoothing does not improve the plane wave expansion method for the 1D TM Mode Problem.

Now, let us consider the sampling method. This method is applied in a similar way

as in Section 4.4. Again we modify $[\gamma]_j$ and $[\log n^2]_j$ from the definition of A in (5.8). We replace $[\gamma]_j$ and $[\log n^2]_j$ with $[\mathrm{Q}_M\,\gamma]_j$ and $[\mathrm{Q}_M \log n^2]_j$ respectively, where $M \in \mathbb{N}$ is fixed and Q_M is the Interpolation Projection defined in Subsection 3.2.5. In Figure 5-7 we have plotted the errors of the eigenvalues and eigenfunctions for the plane wave expansion method with sampling for fixed G ($G = 2^{16} - 1$) and varying grid spacing (varying M). Again, the reference solution is the solution to (5.8) with $G = 2^{18} - 1$ (and exact Fourier coefficients). We see that both the eigenvalue and eigenfunction errors appear to have $\mathcal{O}(M^{-3/2})$ dependence on M. However, $\mathcal{O}(M^{-3/2})$ convergence only appears in a small range of M values (when $M \approx N_f$) for the eigenfunction errors. For $M \gg N_f$, the eigenfunction error does not converge, but this is because the accuracy of the reference solution has been reached (see Figure 5-2). Recall that for the 1D TE Mode Problem we observed $\mathcal{O}(M^{-1})$ error dependence for both the eigenfunction and eigenvalue errors in general but Model Problem 1 was a special case. We are still unsure as to whether or not Model Problem 1 is a special case for the 1D TM Mode Problem and we do not use the results in Figure 5-7 to predict how to choose the grid-spacing in the sampling grid to recover the convergence rate of exact Fourier coefficients.

In Figures 5-8 and 5-9 we have plotted the errors of the plane wave expansion method with sampling for varying G where we have chosen $M = N_f^r$ for different constants r (recall that $N_f = 4G + 4$). Again, the reference solution is the solution to (5.8) with $G = 2^{18} - 1$, i.e. the plane wave expansion method with exact Fourier coefficients. From these plots we observe that if $M \geq N_f^{3/2}$ then we recover the error convergence rate for both the eigenfunctions and eigenvalues of the plane wave expansion method with exact Fourier coefficients, and choosing $M = N_f$ gives us a method that does not converge. Recall that for the 1D TE Mode Problem in Chapter 4 we needed to choose $M \geq N_f^{3/2}$ to recover the $\mathcal{O}(G^{-3/2})$ convergence rate for the eigenfunction error and $M \geq N_f^3$ to recover the $\mathcal{O}(G^{-3})$ convergence rate for the eigenvalue error. If we compare these results then it suggests that the sampling method performs better for the eigenvalue error of the 1D TM Mode Problem than it does for the 1D TE Mode Problem in the sense that a smaller M may be chosen to recover the convergence rate of the plane wave expansion method with exact Fourier coefficients. However, we must temper this "favourable" result by remembering that with $M = N_f^{3/2}$ the eigenvalue errors for the 1D TE Mode Problem will still decay faster ($\mathcal{O}(G^{-3/2})$ vs. $\mathcal{O}(G^{-1})$) than the 1D TM Mode Problem.

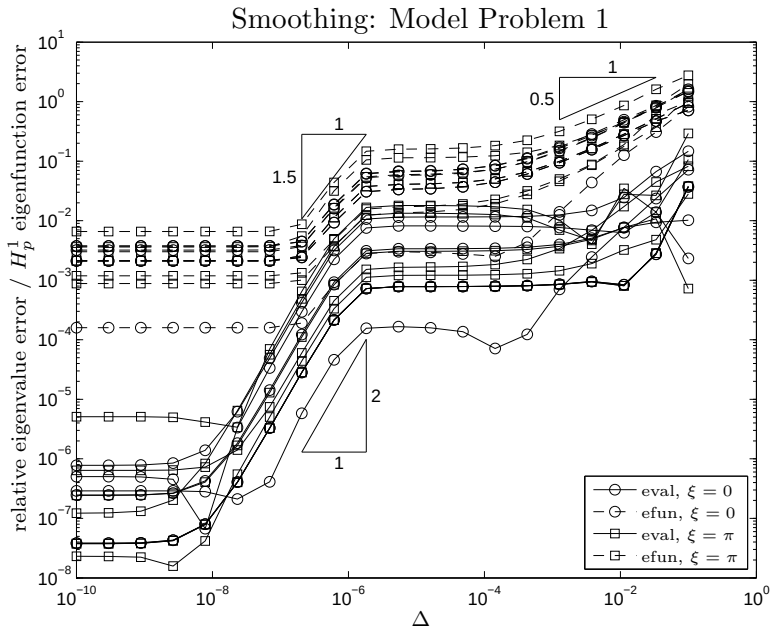

Figure 5-4: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. Δ for the first 5 eigenpairs of the plane wave expansion method with smoothing (fixed G) applied to Model Problem 1 for $\xi = 0$ and $\xi = \pi$.

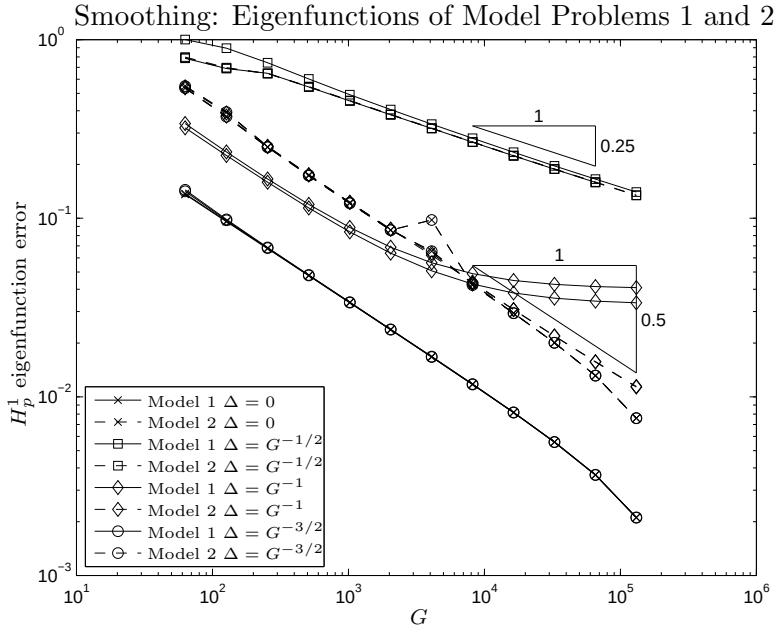

Figure 5-5: Plot of the H_p^1 norm of the error vs. G for the 1st eigenfunction of the plane wave expansion method with smoothing approximation to Problem 5.1 for $\xi = 0$, and $\xi = \pi$ (for Model Problem 1) or $\xi = \frac{\pi}{13}$ (for Model Problem 2).

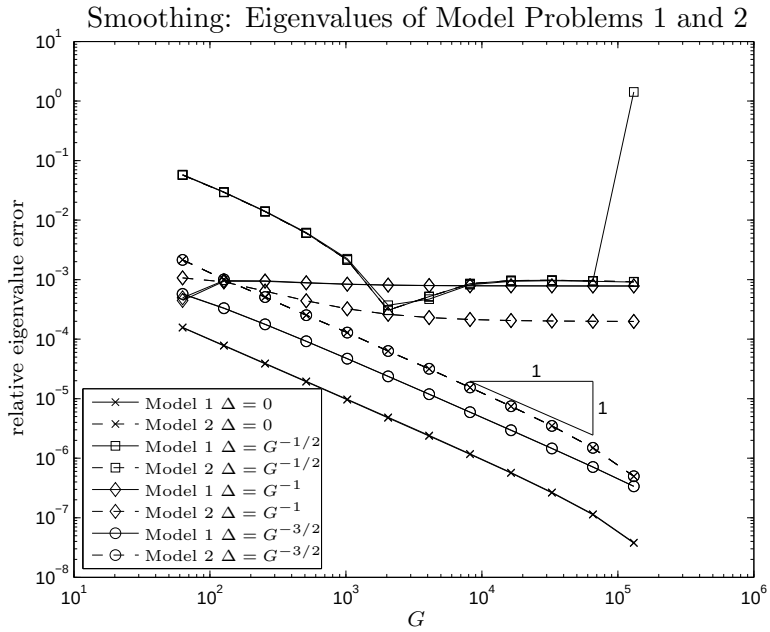

Figure 5-6: Plot of the relative error of the 1st eigenvalue vs. G for the plane wave expansion method with smoothing approximation to Problem 5.1 for $\xi = 0$, and $\xi = \pi$ (for Model Problem 1) or $\xi = \frac{\pi}{13}$ (for Model Problem 2).

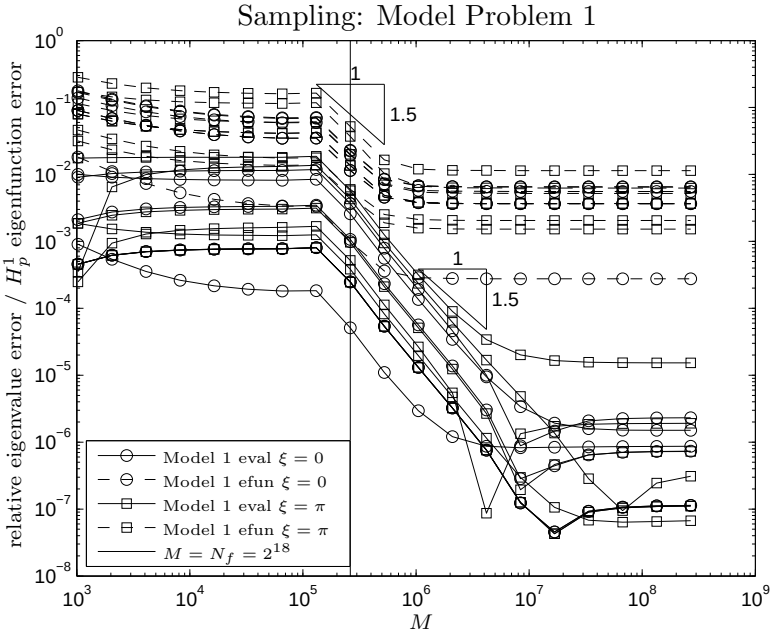

Figure 5-7: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. M for the first 5 eigenpairs of plane wave expansion method with sampling (fixed $G = 2^{16} - 1 \approx 6.5 \times 10^4$) applied to Model Problem 1 for $\xi = 0$, and $\xi = \pi$. $N_f = 2^{18} \approx 2.6 \times 10^5$.

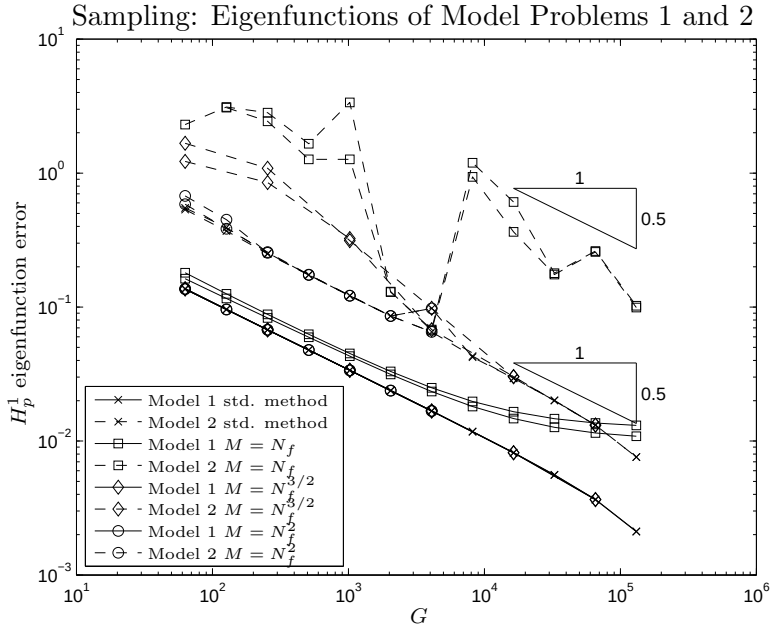

Figure 5-8: Plot of the 1st eigenfunction error vs. G for the plane wave expansion method with sampling applied to Model Problems 1 and 2 where $M = N_f^r$ for different r.

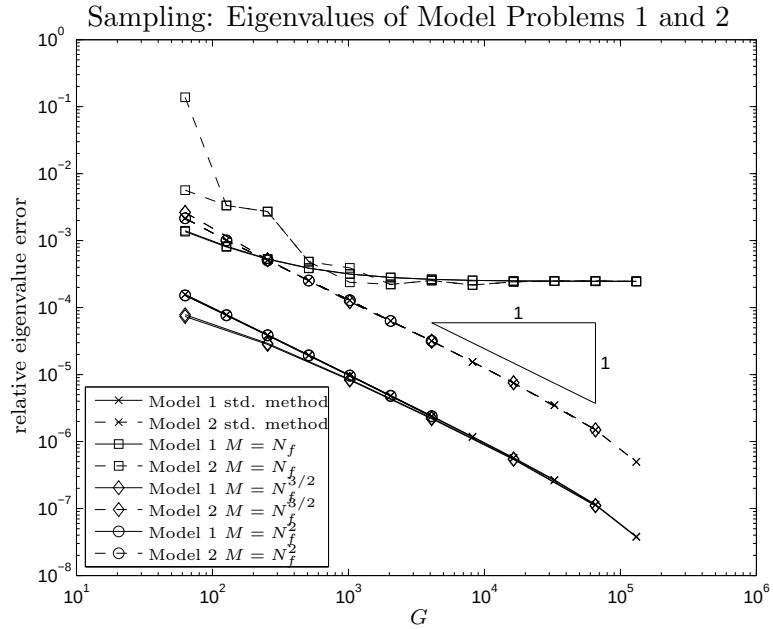

Figure 5-9: Plot of the 1st eigenvalue error vs. G for the plane wave expansion method with sampling applied to Model Problems 1 and 2 where $M = N_f^r$ for different r.

CHAPTER 6
FULL 2D PROBLEM

In this chapter we consider the plane wave expansion method applied to the Full 2D Problem (see Problem 2.1 in Section 2.5). As for the 1D TM Mode Problem (see previous chapter) the error analysis is not as straight forward as for the Scalar 2D Problem or the 1D TE Mode Problem (see Chapter 4). However, unlike the 1D TM Mode Problem, we can not even write the problem in *divergence form* and to gain any insight into the theoretical properties of the problem we will have to consider Maxwell's equations in 3D.

We begin by presenting the plane wave expansion method in the same way as it is done in [64], and we explain how the Fast Fourier Transform is used to obtain an efficient implementation of the method. We also discuss a preconditioner that can be used with the implementation of the plane wave expansion method.

Once we have presented the method that we use we will consider the theoretical analysis of our method. Although we have been unsuccessful in developing a stability result for the plane wave expansion method applied to this problem, we have managed to prove existence of eigenpairs for the exact problem and regularity results for at least some of the eigenfunctions of the exact problem. Since we can not write down the Full 2D Problem in divergence form (as we could for the 1D TM Mode Problem, see (5.2)) we resort to studying Maxwell's equations in 3D. Via Maxwell's equations in 3D we prove that there exist eigenpairs of the Full 2D Problem that are in $H_p^{3/2-\epsilon}$ for some $0 \leq \epsilon < 1/2$. Unfortunately, we can not be sure that all eigenfunctions of the Full 2D Problem share this regularity. Also, recall that for the 1D TM Mode Problem we showed that the eigenfunctions are in $H_p^{3/2-\epsilon}$ for arbitrarily small $\epsilon > 0$. Our result in this chapter is not quite as strong as the result for the 1D TM Mode Problem, but we have not ruled out the possibility that the eigenfunctions of the Full 2D Problem could be in $H_p^{3/2-\epsilon}$ for arbitrarily small $\epsilon > 0$, and we have at least shown that some of the eigenfunctions are in H_p^{1+s} for some $s > 0$.

The regularity result falls short of what we managed to prove for the Scalar 2D Problem in Chapter 4, where we showed the eigenfunctions of the Scalar 2D Problem are in $H_p^{5/2-\epsilon}$ for all $\epsilon > 0$. This deficiency in regularity can be explained by the presence of the additional *vector* or *coupling* term in the equation for the Full 2D Problem (that was not present in the Scalar 2D Problem).

Following our analysis we compute some numerical examples of the plane wave expansion method applied to the Full 2D Problem. In our computations we observe that the eigenvalue errors and the eigenfunction errors decay at the same rates as the 1D TM Mode Problem. That is, we observe that the eigenfunction error decays at the same rate as the approximation error for a function in $H_p^{3/2-\epsilon}$ for arbitrarily small $\epsilon > 0$ approximated by plane waves. This suggests that the eigenfunctions of the Full 2D Problem are in fact in $H_p^{3/2-\epsilon}$ for arbitrarily small $\epsilon > 0$ and that the plane wave expansion method is stable. We also observe that the eigenvalue error decays at twice the rate of the eigenfunction error. This suggests that the problem has a certain degree of symmetry even though the matrix eigenproblem from the plane wave expansion method is non-symmetric. The convergence rates that we observe are not a surprise because, in a certain sense, the Full 2D Problem is the 2D extension of the 1D TM Mode Problem.

Finally, we briefly present a few numerical computations that experiment with the use of the smoothing and sampling methods applied to the Full 2D Problem, and find that with appropriate choices of the smoothing and sampling parameters, we can recover the convergence rates of the standard plane wave expansion method. As for all of the other problems we have examined in previous chapters we find that we can not improve the standard plane wave expansion method by smoothing or sampling.

6.1 The Problem

Unlike the problems we have looked at so far in this thesis, the Full 2D problem is a vectorial problem. Formally, the Full 2D Problem (see Problem 2.1 in Section 2.5) is

$$(\nabla_t^2 + \gamma)\mathbf{h}_t - (\nabla_t \times \mathbf{h}_t) \times (\nabla_t \eta) = \beta^2 \mathbf{h}_t \qquad \text{on } \mathbb{R}^2 \tag{6.1}$$

where $\nabla_t = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, 0)$ and $\mathbf{h}_t = (h_x, h_y, 0)$ is a 2D vector field eigenfunction with components h_x and h_y. The coefficients $\gamma = \gamma(x, y)$ and $\eta = \eta(x, y)$ are piecewise constant, 2D-periodic scalar fields, and β^2 is an eigenvalue. Note that for notational convenience we will keep working with 3D vectors (even though the last component will be 0). In physical terms, $\mathbf{h}_t$ and β both represent different parts of the magnetic field in the following way,

$$\mathbf{H}(\mathbf{x}) = (\mathbf{h}_t(x, y) + h_z(x, y)\hat{\mathbf{z}})\, \mathrm{e}^{i\beta z}\,. \tag{6.2}$$

The z-component of the magnetic field, $h_z(x, y)$, and the electric field are uniquely determined given $\mathbf{h}_t$ and β. See Subsection 2.2.2 for more details on this.

The functions $\gamma(x, y)$ and $\eta(x, y)$ are given by,

$$\gamma(x, y) = \frac{4\pi^2 n^2(x, y)}{\lambda_0^2}$$
$$\eta(x, y) = \log(n^2(x, y))$$

where $n^2(x, y)$ is the refractive index of the photonic crystal or photonic crystal fibre. We assume that the scalar field $n^2(x, y)$ is independent of z (i.e. a genuine 2D scalar field) and that it belongs to our special class of 2D-periodic, piecewise constant functions that we defined in Definition 3.36, with period cell $\Omega = [-\frac{1}{2}, \frac{1}{2}]^2$ and $1 \leq n^2(x, y) \leq n^2_{max}$. Recall that for photonic crystal fibres $n(x, y)$ is not necessarily periodic but we have forced $n^2(x, y)$ to be periodic by applying the supercell method and we are already satisfied that the supercell method converges as the size of the supercell increases. The constant λ_0 specifies the wavelength of light relative to the size of the structure and $\log(\cdot)$ is the natural logarithm.

Notice that (6.1) differs from (4.1) (the equation for the Scalar 2D Problem) only because of the presencse of the $(\nabla_t \times \mathbf{h}_t) \times (\nabla_t \eta(x, y))$ term. In physics literature this term is sometimes referred to as the *vector* or *coupling* term. We can also think of (6.1) as being similar to the equation for the 1D TM Mode Problem, (5.1). The terms of (6.1) are the same as (5.1) in that we have a Schrödinger operator where the potential term is periodic and piecewise constant, with an additional first order term that has a coefficient that is the derivative of a periodic piecewise constant coefficient. A difference between the two equations is that (6.1) is a 2D vector equation while (5.1) is a 1D scalar equation. Another difference from the 1D TM Mode Problem is that we were able to write the 1D TM Mode Problem equation in "divergence form" (see (5.2)), and in doing so we were able to avoid writing a governing equation (or a variational form) with a distribution as a coefficient. Unfortunately, we can not do this for (6.1). The analysis of the 1D TM Mode Problem depended on being able to write the problem in divergence form. Therefore, we can not use the same approach to study the Full 2D Problem as we did for the 1D TM Mode Problem.

In fact, we are not aware of any attempt in the mathematical literature that tackles the Full 2D Problem in a spectral theory framework. However, there are a number of papers in the phyisics literature (from the Centre for Photonics and Photonic Materials in the Physics Department at the University of Bath) that tackle (6.1) from a computational perspective. See for example, [7], [62], [63], [64] and [66].

Without the proper mathematical analysis we proceed as in [39] and assume a certain form for $\mathbf{h}_t$ (the physics literature often refers to this as Bloch theory) to reduce (6.1) to a problem where the eigenfunctions are periodic with period cell Ω.

Note that in the following, since we are not considering the spectrum of an operator on a Hilbert space, we use the term "eigenfunction" for a function that satisfies the governing equation in the distributional sense and we are not referring to eigenfunctions as we defined them in Subsection 3.4.2. The symmetry argument in [39] is as follows: Since $n^2(x,y)$ is periodic in the directions of the lattice vectors (i.e. in the x and y coordinate directions for how we have defined $n^2(x,y)$), it suffices to only consider eigenfunctions of (6.1) that can be written as

$$\mathbf{h}_t(x,y) = \mathrm{e}^{i\boldsymbol{\xi}\cdot\mathbf{x}}\,\mathbf{u}(x,y) \qquad \forall \mathbf{x} \in \mathbb{R}^3 \tag{6.3}$$

where $\boldsymbol{\xi} \in B = [-\pi,\pi]^2 \times \{0\}$ where $\mathbf{u} = (u_1, u_2, 0)$ is a periodic vector field on $\mathbb{R}^2$ with period cell Ω. More general eigenfunctions can then be obtained by taking linear combinations of eigenfunctions with this form. With this expansion of $\mathbf{h}_t$, (6.1) reduces to the following family of eigenproblems, where $\mathbf{u}$ is the new eigenfunction:

$$(\nabla_t + i\boldsymbol{\xi})^2\mathbf{u} + \gamma(x,y)\mathbf{u} - ((\nabla_t + i\boldsymbol{\xi}) \times \mathbf{u}) \times (\nabla_t \eta(x,y)) = \beta^2\mathbf{u} \qquad \text{on } \mathbb{R}^2, \tag{6.4}$$

for $\boldsymbol{\xi} \in B$. Moreover, we can see that given an eigenpair $(\beta^2, \mathbf{u})$ of (6.4) for $\boldsymbol{\xi} \in B$, then $(\beta^2, \mathrm{e}^{i\boldsymbol{\xi}\cdot\mathbf{x}}\,\mathbf{u}(x,y))$ is an eigenpair of (6.1).

Since $\mathbf{u}$ is periodic with period cell Ω, we can now consider the problem of solving (6.4) on Ω with periodic boundary conditions.

6.2 Method and Implementation

In this section we apply the plane wave expansion method to (6.4) for a fixed $\boldsymbol{\xi} \in B$ to obtain a matrix eigenvalue problem. We then give some details for how we solve this matrix eigenvalue problem. We want to solve (6.4) for periodic eigenfunctions $\mathbf{u}$ and eigenvalues $\lambda := \beta^2$.

To help us understand the implementation let us write (6.4) component-wise,

$$(\nabla_t + i\boldsymbol{\xi})^2 u_1 + \gamma u_1 + \frac{\partial \eta}{\partial y}\left((\tfrac{\partial}{\partial x} + i\xi_1)u_2 - (\tfrac{\partial}{\partial y} + i\xi_2)u_1\right) = \lambda u_1 \tag{6.5}$$

$$(\nabla_t + i\boldsymbol{\xi})^2 u_2 + \gamma u_2 - \frac{\partial \eta}{\partial x}\left((\tfrac{\partial}{\partial x} + i\xi_1)u_2 - (\tfrac{\partial}{\partial y} + i\xi_2)u_1\right) = \lambda u_1 \tag{6.6}$$

As in Section 5.2 for the 1D TM mode problem we apply the plane wave expansion method as it is presented in [64], rather than presenting it as a Galerkin method for a variational eigenvalue problem. Since $\mathbf{u}$ in (6.4) is periodic with period cell Ω we can expand u_1 and u_2 in terms of plane waves,

$$u_i(\mathbf{x}) = \sum_{\mathbf{g}\in\mathbb{Z}^2} [u_i]_{\mathbf{g}}\,\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} \qquad \mathbf{x} \in \mathbb{R}^2, i = 1,2.$$

We then substitute this, together with the plane wave expansions of $\gamma(x,y)$ and $\eta(x,y)$ into (6.5) to get

$$\begin{aligned} -\sum_{\mathbf{g}\in\mathbb{Z}^2} |\boldsymbol{\xi}+2\pi\mathbf{g}|^2[u_1]_{\mathbf{g}}\,\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} + \sum_{\mathbf{g}\in\mathbb{Z}^2}\sum_{\mathbf{k}\in\mathbb{Z}^2}[\gamma]_{\mathbf{k}}[u_1]_{\mathbf{g}}\,\mathrm{e}^{i2\pi(\mathbf{k}+\mathbf{g})\cdot\mathbf{x}} \\ -\sum_{\mathbf{g}\in\mathbb{Z}^2}\sum_{\mathbf{k}\in\mathbb{Z}^2}(2\pi k_2)[\eta]_{\mathbf{k}}\Big((\xi_1+2\pi g_1)[u_2]_{\mathbf{g}}-(\xi_2+2\pi g_2)[u_1]_{\mathbf{g}}\Big)\,\mathrm{e}^{i2\pi(\mathbf{k}+\mathbf{g})\cdot\mathbf{x}} \\ = \lambda\sum_{\mathbf{g}\in\mathbb{Z}^2}[u_1]_{\mathbf{g}}\,\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} \qquad \mathbf{x}\in\mathbb{R}^2 \end{aligned} \tag{6.7}$$

and into (6.6) to get

$$\begin{aligned} -\sum_{\mathbf{g}\in\mathbb{Z}^2} |\boldsymbol{\xi}+2\pi\mathbf{g}|^2[u_2]_{\mathbf{g}}\,\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} + \sum_{\mathbf{g}\in\mathbb{Z}^2}\sum_{\mathbf{k}\in\mathbb{Z}^2}[\gamma]_{\mathbf{k}}[u_2]_{\mathbf{g}}\,\mathrm{e}^{i2\pi(\mathbf{k}+\mathbf{g})\cdot\mathbf{x}} \\ +\sum_{\mathbf{g}\in\mathbb{Z}^2}\sum_{\mathbf{k}\in\mathbb{Z}^2}(2\pi k_1)[\eta]_{\mathbf{k}}\Big((\xi_1+2\pi g_1)[u_2]_{\mathbf{g}}-(\xi_2+2\pi g_2)[u_1]_{\mathbf{g}}\Big)\,\mathrm{e}^{i2\pi(\mathbf{k}+\mathbf{g})\cdot\mathbf{x}} \\ = \lambda\sum_{\mathbf{g}\in\mathbb{Z}^2}[u_2]_{\mathbf{g}}\,\mathrm{e}^{i2\pi\mathbf{g}\cdot\mathbf{x}} \qquad \mathbf{x}\in\mathbb{R}^2. \end{aligned} \tag{6.8}$$

Now we multiply (6.7) and (6.8) by $\mathrm{e}^{-i2\pi\mathbf{g}'\cdot\mathbf{x}}$ for $\mathbf{g}'\in\mathbb{Z}^2$ and integrate over Ω to get

$$\sum_{\mathbf{g}\in\mathbb{Z}}\begin{pmatrix}\mathcal{A}_{11} & \mathcal{A}_{12}\\ \mathcal{A}_{21} & \mathcal{A}_{22}\end{pmatrix}\begin{pmatrix}[u_1]_{\mathbf{g}}\\ [u_2]_{\mathbf{g}}\end{pmatrix} = \lambda\begin{pmatrix}[u_1]_{\mathbf{g}'}\\ [u_2]_{\mathbf{g}'}\end{pmatrix} \qquad \forall\mathbf{g}'\in\mathbb{Z}^2 \tag{6.9}$$

where the $\mathcal{A}_{ij}$ are given by

$$\begin{aligned} \mathcal{A}_{11}(\mathbf{g}',\mathbf{g}) &= -|\boldsymbol{\xi}+2\pi\mathbf{g}|^2\delta_{\mathbf{g},\mathbf{g}'} + [\gamma]_{\mathbf{g}'-\mathbf{g}} + 2\pi(g_2'-g_2)(\xi_2+2\pi g_2)[\eta]_{\mathbf{g}'-\mathbf{g}} \\ \mathcal{A}_{12}(\mathbf{g}',\mathbf{g}) &= -2\pi(g_2'-g_2)(\xi_1+2\pi g_1)[\eta]_{\mathbf{g}'-\mathbf{g}} \\ \mathcal{A}_{21}(\mathbf{g}',\mathbf{g}) &= -2\pi(g_1'-g_1)(\xi_2+2\pi g_2)[\eta]_{\mathbf{g}'-\mathbf{g}} \\ \mathcal{A}_{22}(\mathbf{g}',\mathbf{g}) &= -|\boldsymbol{\xi}+2\pi\mathbf{g}|^2\delta_{\mathbf{g},\mathbf{g}'} + [\gamma]_{\mathbf{g}'-\mathbf{g}} + 2\pi(g_1'-g_1)(\xi_1+2\pi g_1)[\eta]_{\mathbf{g}'-\mathbf{g}} \end{aligned} \tag{6.10}$$

To create a finite dimensional problem we restrict $\mathbf{g}$ and $\mathbf{g}'$ so that $|\mathbf{g}|,|\mathbf{g}'|\le G$ for a chosen $G\in\mathbb{N}$. This is equivalent to restricting $\mathbf{g}$ and $\mathbf{g}'$ so that $\mathbf{g},\mathbf{g}'\in\mathbb{Z}^2_{G,\mathrm{o}}$, or $[u_1]_{\mathbf{g}}=[u_2]_{\mathbf{g}}=0$ for all $|\mathbf{g}|>G$. To define a matrix eigenproblem that is equivalent to the finite dimensional problem we first define $N:=\dim\mathbb{Z}^2_{G,\mathrm{o}}$ and a one-to-one map $i:\mathbb{Z}^2_{G,\mathrm{o}}\to\{n\in\mathbb{N}:n\le N\}$ that orders the elements in $\mathbb{Z}^2_{G,\mathrm{o}}$ in ascending order, i.e. $i(\mathbf{g})<i(\mathbf{g}')$ if $|\mathbf{g}|<|\mathbf{g}'|$. The $2N\times 2N$ matrix eigenproblem is then

$$\mathrm{A}\,\mathbf{x} = \lambda_G\mathbf{x} \tag{6.11}$$

where A and $\mathbf{x}$ can be split into $N \times N$ submatrices and subvectors of length N,

$$\mathrm{A} = \begin{bmatrix} \mathrm{A}_{11} & \mathrm{A}_{12} \\ \mathrm{A}_{21} & \mathrm{A}_{22} \end{bmatrix} \qquad \mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$$

and the submatricies and subvectors have entries defined by (see (6.10))

$$\begin{aligned} (\mathrm{A}_{11})_{i(\mathbf{g}'),i(\mathbf{g})} &= \mathcal{A}_{11}(\mathbf{g}', \mathbf{g}) \\ (\mathrm{A}_{12})_{i(\mathbf{g}'),i(\mathbf{g})} &= \mathcal{A}_{12}(\mathbf{g}', \mathbf{g}) \\ (\mathrm{A}_{21})_{i(\mathbf{g}'),i(\mathbf{g})} &= \mathcal{A}_{21}(\mathbf{g}', \mathbf{g}) \\ (\mathrm{A}_{22})_{i(\mathbf{g}'),i(\mathbf{g})} &= \mathcal{A}_{22}(\mathbf{g}', \mathbf{g}) \qquad \forall \mathbf{g}, \mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}} \end{aligned}$$

and

$$\begin{aligned} (\mathbf{x}_1)_{i(\mathbf{g})} &= [u_1]_{\mathbf{g}} \\ (\mathbf{x}_2)_{i(\mathbf{g})} &= [u_2]_{\mathbf{g}} \qquad \mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}. \end{aligned} \tag{6.12}$$

To solve (6.11) we use the same implementation and a similar preconditioner that we have used throughout this thesis. Namely, we use an iterative eigensolver (Implicitly Restarted Arnoldi method) since we are only interested in a small number of extremal eigenvalues of (6.11). We apply our eigensolver to A^{-1} (instead of A because this gives us better convergence towards the smallest eigenvalues of A) and at each iteration of the eigensolver we are required to solve a linear system to obtain the operation of A^{-1}. We use GMRES to do this because A is non-symmetric. In the inner iteration of the GMRES algorithm we are required to compute matrix-vector products with A. Since A is in general very large and dense, the efficiency of the method for solving (6.11) depends crucially on our ability to compute $\mathrm{A}\,\mathbf{v}$ efficiently. We obtain such an efficient algorithm for computing $\mathrm{A}\,\mathbf{v}$ by taking advantage of the submatrix structure of A. With $\mathbf{v}$ split into two subvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ of length N as in (6.12) we can reduce the problem of computing $\mathrm{A}\,\mathbf{v}$ efficiently to the problem of computing $\mathrm{A}_{11}\,\mathbf{v}_1$, $\mathrm{A}_{12}\,\mathbf{v}_2$, $\mathrm{A}_{21}\,\mathbf{v}_1$ and $\mathrm{A}_{22}\,\mathbf{v}_2$ efficiently.

From (6.10) we realise that each of the submatrices A_{ij} can be expanded in the following way,

$$\begin{aligned} \mathrm{A}_{11} &= -\,\mathrm{D} + \mathrm{V} + \mathrm{W}_2\,\mathrm{D}_2 \\ \mathrm{A}_{12} &= -\,\mathrm{W}_2\,\mathrm{D}_1 \\ \mathrm{A}_{21} &= -\,\mathrm{W}_1\,\mathrm{D}_2 \\ \mathrm{A}_{22} &= -\,\mathrm{D} + \mathrm{V} + \mathrm{W}_1\,\mathrm{D}_1 \end{aligned}$$

where D, D_1 and D_2 are all diagonal matrices with entries given by

$$\begin{aligned}
\mathrm{D}_{i(\mathbf{g}),i(\mathbf{g})} &= |\boldsymbol{\xi} + 2\pi\mathbf{g}|^2 \\
(\mathrm{D}_1)_{i(\mathbf{g}),i(\mathbf{g})} &= \xi_1 + 2\pi g_1 \\
(\mathrm{D}_2)_{i(\mathbf{g}),i(\mathbf{g})} &= \xi_2 + 2\pi g_2 \qquad \forall \mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}
\end{aligned}$$

and V, W_1 and W_2 are dense matrices with entries given by

$$\begin{aligned}
\mathrm{V}_{i(\mathbf{g}'),i(\mathbf{g})} &= [\gamma]_{\mathbf{g}'-\mathbf{g}} \\
\mathrm{W}_{i(\mathbf{g}'),i(\mathbf{g})} &= 2\pi(g_1' - g_1)[\log n^2]_{\mathbf{g}'-\mathbf{g}} = [i\tfrac{\partial}{\partial x}(\log n^2)]_{\mathbf{g}'-\mathbf{g}} \\
\mathrm{W}_{i(\mathbf{g}'),i(\mathbf{g})} &= 2\pi(g_2' - g_2)[\log n^2]_{\mathbf{g}'-\mathbf{g}} = [i\tfrac{\partial}{\partial y}(\log n^2)]_{\mathbf{g}'-\mathbf{g}} \qquad \forall \mathbf{g}, \mathbf{g}' \in \mathbb{Z}^2_{G,\mathrm{o}}.
\end{aligned}$$

Obviously, it is very cheap to compute matrix-vector products with D, D_1 and D_2 because they are diagonal matrices. To compute matrix-vector products with V, W_1 and W_2 we use a similar algorithm to Algorithm 4.19, each at a cost of $\mathcal{O}(N \log N)$ operations. From our work so far it appears that to compute $\mathrm{A}\,\mathbf{v}$ will require 12 FFTs or inverse FFTs (two applications of V, W_1 and W_2 requiring two FFTs each). In actual fact, we can reduce this number to 6 (see Algorithm 6.1 below).

For completeness, we now present the complete algorithm for computing $\mathrm{A}\,\mathbf{v}$ for a given vector $\mathbf{v} \in \mathbb{C}^{2N}$. As in Chapter 4 we choose $N_f = 2^n$ for $n \in \mathbb{N}$ (to get the best performance for our FFT), set $G = \frac{N_f}{4} - 1$, then $N = \dim \mathbb{Z}^2_{G,\mathrm{o}}$. We also use the same matrix notation convention that we used in Chapter 4 (see just before Algorithm 4.19) where $X, Y, \widehat{X}$ and $\widehat{Y}$ represent functions in $\mathcal{T}^2_{N_f}$ with nodal values (X and Y) or Fourier coefficients ($\widehat{X}$ and $\widehat{Y}$), so that for example, $\widehat{X} = \mathrm{fft}(X)$ and $X = \mathrm{ifft}(\widehat{X})$. Let $\mathbf{g}_0 := (\frac{N_f}{2} + 1, \frac{N_f}{2} + 1) = (2G + 3, 2G + 3)$.

Algorithm 6.1. Let $\mathbf{v} \in \mathbb{C}^{2N}$, let $\widehat{Y}_1$ be a matrix of Fourier coefficients of $(i\frac{\partial}{\partial x}(\log n^2))$, let $\widehat{Y}_2$ be a matrix of Fourier coefficients of $(i\frac{\partial}{\partial y}(\log n^2))$ and let $\widehat{Z}$ be a matrix of Fourier coefficients of γ, so that

$$\begin{aligned}
(\widehat{Y}_1)_{ij} &= (2\pi g_1)[\log n^2]_{\mathbf{g}} \\
(\widehat{Y}_2)_{ij} &= (2\pi g_2)[\log n^2]_{\mathbf{g}} \\
(\widehat{Z})_{ij} &= [\gamma]_{\mathbf{g}}
\end{aligned}$$

where $\mathbf{g} = (i, j) - \mathbf{g}_0$ and $i, j = 1, \ldots, N_f$. Pre-compute $Y_1 \leftarrow \mathrm{ifft}(\widehat{Y})$, $Y_2 \leftarrow \mathrm{ifft}(\widehat{Y}_2)$ and $Z \leftarrow \mathrm{ifft}(\widehat{Z})$ and compute $\mathrm{A}\,\mathbf{v}$ in the following way.
$\widehat{V}_1, \widehat{V}_2, \widehat{A}_1, \widehat{A}_2, \widehat{B}_1, \widehat{B}_1 \leftarrow 0$.
$(\widehat{V}_1)_{\mathbf{g}+\mathbf{g}_0} \leftarrow \mathbf{v}_{i(\mathbf{g})}$ for $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.
$(\widehat{V}_2)_{\mathbf{g}+\mathbf{g}_0} \leftarrow \mathbf{v}_{i(\mathbf{g})+N}$ for $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.
$(\widehat{A}_1)_{\mathbf{g}+\mathbf{g}_0} \leftarrow |\boldsymbol{\xi} + 2\pi\mathbf{g}|^2 (\widehat{V}_1)_{\mathbf{g}+\mathbf{g}_0}$ for $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.

$(\widehat{A}_2)_{\mathbf{g}+\mathbf{g}_0} \leftarrow |\boldsymbol{\xi} + 2\pi\mathbf{g}|^2(\widehat{V}_2)_{\mathbf{g}+\mathbf{g}_0}$ for $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.
$(\widehat{B}_1)_{\mathbf{g}+\mathbf{g}_0} \leftarrow (\xi_2 + 2\pi g_2)(\widehat{V}_1)_{\mathbf{g}+\mathbf{g}_0}$ for $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.
$(\widehat{B}_2)_{\mathbf{g}+\mathbf{g}_0} \leftarrow (\xi_1 + 2\pi g_1)(\widehat{V}_2)_{\mathbf{g}+\mathbf{g}_0}$ for $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.
$V_1 \leftarrow \mathrm{ifft}(\widehat{V}_1)$.
$V_2 \leftarrow \mathrm{ifft}(\widehat{V}_2)$.
$B_1 \leftarrow \mathrm{ifft}(\widehat{B}_1)$.
$B_2 \leftarrow \mathrm{ifft}(\widehat{B}_2)$.
$(V_1)_{ij} \leftarrow (Z)_{ij}(V_1)_{ij} + (Y_2)_{ij}(B_1)_{ij} - (Y_2)_{ij}(B_2)_{ij}$ for $i, j = 1, \ldots, N_f$.
$(V_2)_{ij} \leftarrow (Z)_{ij}(V_2)_{ij} + (Y_1)_{ij}(B_2)_{ij} - (Y_1)_{ij}(B_1)_{ij}$ for $i, j = 1, \ldots, N_f$.
$\widehat{V}_1 \leftarrow \mathrm{fft}(V_1)$.
$\widehat{V}_2 \leftarrow \mathrm{fft}(V_2)$.
$\widehat{V}_1 \leftarrow \widehat{V}_1 - \widehat{A}_1$.
$\widehat{V}_2 \leftarrow \widehat{V}_2 - \widehat{A}_2$.
$(\mathrm{A}\,\mathbf{v})_{i(\mathbf{g})} \leftarrow (\widehat{V}_1)_{\mathbf{g}+\mathbf{g}_0}$ for $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.
$(\mathrm{A}\,\mathbf{v})_{i(\mathbf{g})+N} \leftarrow (\widehat{V}_2)_{\mathbf{g}+\mathbf{g}_0}$ for $\mathbf{g} \in \mathbb{Z}^2_{G,\mathrm{o}}$.

We see that Algorithm 6.1 we require only 2 FFTs and 4 inverse FFTs. The total cost of Algorithm 6.1 is $\mathcal{O}(N \log N)$.

To precondition the coefficient matrix A when we solve linear systems we use a similar preconditioner that we have used in the previous chapters. We use

$$\mathrm{P} = \begin{bmatrix} \mathrm{P}_{11} & \mathrm{P}_{12} \\ \mathrm{P}_{21} & \mathrm{P}_{22} \end{bmatrix}$$

where P_{ij} are $N \times N$ submatrices defined as

$$\mathrm{P}_{11} = \begin{bmatrix} B_{11} & 0 \\ 0 & D_{11} \end{bmatrix} \qquad \mathrm{P}_{12} = \begin{bmatrix} B_{12} & 0 \\ 0 & 0 \end{bmatrix}$$
$$\mathrm{P}_{21} = \begin{bmatrix} B_{21} & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathrm{P}_{22} = \begin{bmatrix} B_{22} & 0 \\ 0 & D_{22} \end{bmatrix}$$

where the matrices B_{ij} are $N_b \times N_b$ dense matrices and D_{ii} are $(N - N_b) \times (N - N_b)$ diagonal matrices defined by

$$(B_{ij})_{k\ell} = (\mathrm{A}_{ij})_{k\ell} \qquad \text{for } i, j = 1, 2 \text{ and } k, \ell = 1, \ldots N_b$$
$$(D_{ii})_{kk} = (\mathrm{A}_{ii})_{kk} \qquad \text{for } i = 1, 2 \text{ and } k = 1, \ldots, (N - N_b).$$

In practice we can choose N_b up to 1000.

Although we do not have a theoretical result to prove it, we observe that as in the case of the Scalar 2D Problem in Chapter 4 this preconditioner is optimal in the sense that the number of iterations required by the GMRES algorithm does not appear to

depend on N.

Finally, we write down a discrete variational eigenproblem that is equivalent to the plane wave expansion method and (6.11). For the error analysis of the plane wave expansion method applied to the Full 2D Problem we would like to know how this problem approximates (6.4).

Problem 6.2. For $G \in \mathbb{N}$ find λ_G and $0 \neq \mathbf{u} \in (\mathcal{S}_G)^2$ such that

$$\begin{aligned} a_1(\mathbf{u}, \mathbf{v}) &= \lambda_G b_1(\mathbf{u}, \mathbf{v}) \\ a_2(\mathbf{u}, \mathbf{v}) &= \lambda_G b_2(\mathbf{u}, \mathbf{v}) \qquad \forall \mathbf{v} \in (\mathcal{S}_G)^2 \end{aligned}$$

where

$$\begin{aligned} a_1(\mathbf{u}, \mathbf{v}) &= \int_\Omega (\nabla_t + i\boldsymbol{\xi})^2 u_1 \overline{v_1} + \gamma u_1 \overline{v_1} + \tfrac{\partial \eta}{\partial y}((\tfrac{\partial}{\partial x} + i\xi_1)u_2 - (\tfrac{\partial}{\partial y} + i\xi_2)u_1)\overline{v_1} dx \\ a_2(\mathbf{u}, \mathbf{v}) &= \int_\Omega (\nabla_t + i\boldsymbol{\xi})^2 u_2 \overline{v_2} + \gamma u_2 \overline{v_2} - \tfrac{\partial \eta}{\partial x}((\tfrac{\partial}{\partial x} + i\xi_1)u_2 - (\tfrac{\partial}{\partial y} + i\xi_2)u_1)\overline{v_1} dx \\ b_1(\mathbf{u}, \mathbf{v}) &= \int_\Omega u_1 \overline{v_1} dx \\ b_2(\mathbf{u}, \mathbf{v}) &= \int_\Omega u_2 \overline{v_2} dx. \end{aligned}$$

6.3 Regularity and Error Analysis

In this section we discuss our efforts to analyze the Full 2D Problem and the errors of the plane wave expansion method applied to this problem. First, we discuss the difference between the Full 2D Problem and the 1D TM Mode Problem and why we can not use the approach that we used in the previous chapter. Instead, we resort to considering Maxwell's equations in 3D. Using theory developed in [24] we apply Floquet theory to the 3D problem and we write down a 3D variational eigenvalue problem that is related to (6.4). From this variational eigenvalue problem we are then able to confirm the existence of eigenpairs of (6.4) as well as determining a regularity result for at least some of the eigenfunctions of (6.4). Our regularity result allows us to guarantee that the approximation error of plane waves approximating some of the eigenfunctions of (6.4) (measured in the H_p^1 norm) will decay to zero if the number of plane waves increases. If we assume that the plane wave expansion method applied to (6.4) is stable, i.e. the errors are bounded in terms of the approximation error, then the plane wave expansion method will converge. Unfortunately, we have not yet been able to prove this stability result and we have not been able to prove that *all* of the eigenfunctions of (6.4) share the same regularity result.

Unlike the 1D TM Mode problem we could not find a way to write (6.1) in "diver-

gence form" (or "curl form" for that matter), i.e we could not write (6.1) as

$$\nabla \cdot (F(\mathbf{h}_t)) = \beta^2 G(\mathbf{h}_t)$$

or

$$\nabla \times (F(\mathbf{h}_t)) = \beta^2 G(\mathbf{h}_t)$$

where F and G are differential operators with $L^\infty(\mathbb{R}^2)$ coefficients. Therefore, we were not able to follow the approach from Chapter 5 and write down a variational eigenvalue problem, from which it would be possible to determine the regularity of the eigenfunctions. Instead, we have had to find a different way of writing down a variational problem that is equivalent to (6.1) in order to determine the regularity of the eigenfunctions and in order to study the convergence of Problem 6.2 as $G \to \infty$.

The standard approach would be to multiply each component of (6.1) by a test function $\phi \in C^\infty(\mathbb{R}^2)$, integrate over $\mathbb{R}^2$ and take the closure of the subsequent bilinear form with respect to $(C^\infty(\mathbb{R}^2))^2$. Since $\nabla_t \eta$ is not a classical function, it is not clear to us how to do this, in particular how to choose the appropriate Hilbert space, and we do not get a variational problem that is easy to work with. Thus, we had to consider an alternative approach.

Our idea for approaching this problem is to go back to Maxwell's equations in 3D from which (6.1) was derived. It follows from our derivations in Chapter 2 that if $(\beta^2, \mathbf{h}_t)$ is an eigenpair of (6.1) then

$$\mathbf{H}(\mathbf{x}) = (h_x(x,y), h_y(x,y), \tfrac{i}{\beta}\nabla_t \cdot \mathbf{h}_t(x,y))\, \mathrm{e}^{i\beta z} \tag{6.13}$$

must satisfy the time-harmonic 3D Maxwell equations,

$$\begin{aligned} \nabla \times \left(\tfrac{1}{n^2}\nabla \times \mathbf{H}\right) - k_0^2 \mathbf{H} &= 0 \\ \nabla \cdot \mathbf{H} &= 0 \end{aligned} \tag{6.14}$$

on $\mathbb{R}^3$ in the distributional sense (see Subsections 2.2.1 and 2.2.2). Moreover, if we have a solution to (6.14) and $\mathbf{H}$ has the form (6.13) then we must also have an eigenpair of (6.1).

If we think of k_0^2 in (6.14) as an eigenvalue then we can express (6.14) as an operator on a Hilbert space, where the operator is

$$L = \nabla \times \tfrac{1}{n^2}\nabla \times$$

on the Hilbert space $\{f \in (L^2(\mathbb{R}^3))^3 : \nabla \times f \in (L^2(\mathbb{R}^3))^3, \|\nabla \cdot f\|_{L^2_p} = 0\}$.

We then recognise that since $n^2(x,y)$ is periodic with respect to x and y and constant with respect to z, $n^2(x,y)$ is periodic in all three coordinate directions and L is an

operator with periodic coefficients. Following the work in [24], we can apply Floquet theory to this operator to obtain the following family of operators:

$$L_{\mathbf{k}} = (\nabla + i\mathbf{k}) \times \tfrac{1}{n^2}(\nabla + i\mathbf{k})\times$$

for $\mathbf{k} \in Q = [-\pi, \pi]^3$, where each operator operates on the Hilbert space

$$F_{\mathbf{k}} = \{\mathbf{f} \in (L_p^2)^3 : \nabla \times \mathbf{f} \in (L_p^2)^3, \|(\nabla + i\mathbf{k}) \cdot \mathbf{f}\|_{L_p^2} = 0\}.$$

According to [24] $L_{\mathbf{k}}$ has compact resolvent and so $\sigma(L_{\mathbf{k}})$ is discrete. We can also find the following result in [24] that is similar to Theorem 3.63,

$$\sigma(L) = \overline{\bigcup_{\mathbf{k}\in Q} \sigma(L_{\mathbf{k}})}.$$

Since $\sigma(L_{\mathbf{k}})$ is discrete for each $\mathbf{k} \in Q$, we can write down the following variational eigenvalue problem.

Problem 6.3. For $\mathbf{k} \in Q$, find $\lambda \in \mathbb{R}$ and $0 \neq \mathbf{u} \in F_{\mathbf{k}}$ such that

$$a(\mathbf{u}, \mathbf{v}) = \lambda b(\mathbf{u}, \mathbf{v}) \qquad \forall \mathbf{v} \in F_{\mathbf{k}} \tag{6.15}$$

where

$$a(\mathbf{u}, \mathbf{v}) = \int_\Omega \tfrac{1}{n^2}(\nabla + i\mathbf{k}) \times \mathbf{u} \cdot \overline{(\nabla + i\mathbf{k}) \times \mathbf{v}} d\mathbf{x}$$
$$b(\mathbf{u}, \mathbf{v}) = (\mathbf{u}, \mathbf{v})_{(L_p^2)^3} = \int_\Omega \mathbf{u} \cdot \overline{\mathbf{v}} d\mathbf{x}$$

Before we prove the existence of eigenpairs to Problem 6.3 let us make some definitions and examine the properties of the function space $F_{\mathbf{k}}$. Define the following function spaces

$$H_p(curl) = \{\mathbf{f} \in (L_p^2)^3 : \nabla \times \mathbf{f} \in (L_p^2)^3\}$$
$$H_p(div) = \{\mathbf{f} \in (L_p^2)^3 : \nabla \cdot \mathbf{f} \in L_p^2\}$$

and equip them with the following norms,

$$\|\mathbf{f}\|_{H_p(curl)} = \left(\|\mathbf{f}\|^2_{(L_p^2)^3} + \|\nabla \times \mathbf{f}\|^2_{(L_p^2)^3}\right)^{1/2} \qquad \forall \mathbf{f} \in H_p(curl)$$
$$\|\mathbf{f}\|_{H_p(div)} = \left(\|\mathbf{f}\|^2_{(L_p^2)^3} + \|\nabla \cdot \mathbf{f}\|_{L_p^2}\right)^{1/2} \qquad \forall \mathbf{f} \in H_p(div).$$

We equip $F_{\mathbf{k}}$ with the $H_p(curl)$ norm so that $\|\cdot\|_{S_{\mathbf{k}}} = \|\cdot\|_{H_p(curl)}$. We also define the

following function space,

$$G_{\mathbf{k}} = \{\mathbf{f} \in (L_p^2)^3 : \mathbf{f} = (\nabla + i\mathbf{k})g, g \in H_p^1\}$$

With these definitions of function spaces and their norms we can state some well-known properties that $F_{\mathbf{k}}$, $H_p(curl)$, $H_p(div)$ and $G_{\mathbf{k}}$ possess. Note that the symbol "$\subset\subset$" indicates a *compact embedding* (for a definition see page 271 of [21]).

Lemma 6.4. *With $\mathbf{k} \in Q$, we can state the following properties of $F_{\mathbf{k}}$,*

1. $F_{\mathbf{k}} \subset H_p(curl) \cap H_p(div) \subset (H_p^{1/2})^3$.

2. $F_{\mathbf{k}} \subset\subset (L_p^2)^3$.

3. $(H_p^1)^3 \subsetneq H_p(curl)$.

4. $(L_p^2)^3 = F_{\mathbf{k}} \oplus G_{\mathbf{k}}$.

Proof. Part 1. $F_{\mathbf{k}} \subset H_p(curl)$ follows directly from the definition of $F_{\mathbf{k}}$. $F_{\mathbf{k}} \subset H_p(div)$ follows from the fact that $\nabla \cdot \mathbf{f} = -i\mathbf{k} \cdot \mathbf{f}$ and $\mathbf{f} \in (L_p^2)^3$ for all $\mathbf{f} \in F_{\mathbf{k}}$. Therefore $F_{\mathbf{k}} \subset H_p(curl) \cap H_p(div)$. To prove that $H_p(curl) \cap H_p(div) \subset (H_p^{1/2})^3$ we use Theorem 3.47 on page 69 of [57] which states: Let $\widetilde{\Omega} \subset \mathbb{R}^3$ be a bounded Lipschitz domain and let ν defind the outward pointing normal of $\partial\widetilde{\Omega}$. Suppose $\mathbf{u} \in (L^2(\widetilde{\Omega}))^3$ such that $\nabla \times \mathbf{u} \in (L^2(\widetilde{\Omega}))^3$, $\nabla \cdot \mathbf{u} \in L^2(\widetilde{\Omega})$ and $\mathbf{u} \times \boldsymbol{\nu} \in (L^2(\widetilde{\Omega}))^3$. Then $\mathbf{u} \in (H^{1/2}(\widetilde{\Omega}))^3$ and

$$\|\mathbf{u}\|_{(H^{1/2}(\widetilde{\Omega}))^3} \lesssim \|\mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} + \|\nabla \times \mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} + \|\nabla \cdot \mathbf{u}\|_{L^2(\widetilde{\Omega})} + \|\mathbf{u} \times \boldsymbol{\nu}\|_{(L^2(\partial\widetilde{\Omega}))^3}. \quad (6.16)$$

We now show that $H_p(curl) \cap H_p(div) \subset (H_p^{1/2})^3$. Define $\theta \in \mathcal{D}(\mathbb{R}^3)$ and $\widetilde{\Omega}$ as in Lemma 3.17 and let $\mathbf{u} \in H_p(curl) \cap H_p(div)$. Then

$$\begin{aligned}
\|\mathbf{u}\|_{(H_p^{1/2})^3} &\lesssim \|\theta\mathbf{u}\|_{(H^{1/2}(\mathbb{R}^3))^3} && \text{by Theorem 3.29} \\
&= \|\theta\mathbf{u}\|_{(H^{1/2}(\widetilde{\Omega}))^3} && \text{since } \operatorname{supp}\theta \subset \widetilde{\Omega} \\
&\lesssim \|\theta\mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} + \|\nabla \times (\theta\mathbf{u})\|_{(L^2(\widetilde{\Omega}))^3} \\
&\quad + \|\nabla \cdot (\theta\mathbf{u})\|_{L^2(\widetilde{\Omega})} + \|(\theta\mathbf{u}) \times \nu\|_{(\mathbf{L^2}(\partial\widetilde{\mathbf{\Omega}}))^3} && \text{by (6.16)} \\
&= \|\theta\mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} + \|\nabla \times (\theta\mathbf{u})\|_{(L^2(\widetilde{\Omega}))^3} + \|\nabla \cdot (\theta\mathbf{u})\|_{L^2(\widetilde{\Omega})} && \text{since } \theta\mathbf{u}|_{\partial\widetilde{\Omega}} = 0 \\
&\leq \|\theta\mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} + \|\theta\nabla \times \mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} + \|(\nabla\theta) \times \mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} \\
&\quad + \|\theta\nabla \cdot \mathbf{u}\|_{L^2(\widetilde{\Omega})} + \|(\nabla\theta) \cdot \mathbf{u}\|_{L^2(\widetilde{\Omega})}
\end{aligned}$$

Continuing,

$$\begin{aligned}
\|\mathbf{u}\|_{(H_p^{1/2})^3} &\lesssim \|\theta\mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} + \|\theta\nabla\times\mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} + \|\mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} \\
&\quad + \|\theta\nabla\cdot\mathbf{u}\|_{L^2(\widetilde{\Omega})} + \|\mathbf{u}\|_{(L^2(\widetilde{\Omega}))^3} && \text{since } \theta\in\mathcal{D}(\mathbb{R}^3) \\
&\lesssim \|\theta\mathbf{u}\|_{(L^2(\mathbb{R}^3))^3} + \|\theta\nabla\times\mathbf{u}\|_{(L^2(\mathbb{R}^3))^3} + \|\mathbf{u}\|_{(L^2(\Omega))^3} && \text{since } \operatorname{supp}\theta\subset\widetilde{\Omega} \\
&\quad + \|\theta\nabla\cdot\mathbf{u}\|_{L^2(\mathbb{R}^3)} + \|\mathbf{u}\|_{(L^2(\Omega))^3} && \text{and } u \text{ is periodic} \\
&\lesssim \|\mathbf{u}\|_{(L_p^2)^3} + \|\nabla\times\mathbf{u}\|_{(L_p^2)^3} + \|\nabla\cdot\mathbf{u}\|_{L_p^2} && \text{by Theorem 3.29} \\
&\lesssim \|\mathbf{u}\|_{H_p(curl)} + \|\mathbf{u}\|_{H_p(div)}.
\end{aligned}$$

Therefore, $\mathbf{u}\in(H_p^{1/2})^3$ and $H_p(curl)\cap H_p(div)\subset(H_p^{1/2})^3$.

Part 2. The compact embedding $F_{\mathbf{k}}\subset\subset(L_p^2)^3$ follows from the fact that $F_{\mathbf{k}}$ is continuously embedded in $(H_p^{1/2})^3$ (Part 1) and that $H_p^{1/2}\subset\subset L_p^2$ (see Lemma 3.24).

Part 3. It is obvious that $(H_p^1)^3\subset H_p(curl)$ since $\|\nabla\times\mathbf{f}\|_{(L_p^2)^3}\lesssim\|\mathbf{f}\|_{(H_p^1)^3}$ for all $\mathbf{f}\in(H_p^1)^3$. To show that $(H_p^1)^3\neq H_p(curl)$ we can construct a function that is in $H_p(curl)$ but not in $(H_p^1)^3$. For example, a function $\mathbf{u}=(u,0,0)$ with $u\in L_p^2$, $D_{x_2}u\in L_p^2$, $D_{x_3}u\in L_p^2$, but $D_{x_1}u\notin L_p^2$ satisfies $\mathbf{u}\in H_p(curl)$ and $\mathbf{u}\notin(H_p^1)^3$.

Part 4. This result is known as a Helmholtz decomposition and is given in [24]. □

Now let us prove the following lemma about $a(\cdot,\cdot)$ from Problem 6.3.

Lemma 6.5. *The bilinear form $a(\cdot,\cdot)$ from Problem 6.3 is bounded and Hermitian on $F_{\mathbf{k}}$, as well as satisfying*

$$a(\mathbf{v},\mathbf{v}) + \tfrac{6\pi^2+1}{2n_{max}^2}\|\mathbf{v}\|_{(L_p^2)^3}^2 \gtrsim \|\mathbf{v}\|_{S_{\mathbf{k}}}^2 \qquad \forall\mathbf{v}\in F_{\mathbf{k}}. \tag{6.17}$$

Proof. First, let us show that $a(\cdot,\cdot)$ is bounded on $F_{\mathbf{k}}$. For $\mathbf{u},\mathbf{v}\in F_{\mathbf{k}}$ we get,

$$\begin{aligned}
|a(\mathbf{u},\mathbf{v})| &= \left|\int_\Omega \tfrac{1}{n^2}(\nabla+i\mathbf{k})\times\mathbf{u}\cdot\overline{(\nabla+i\mathbf{k})\times\mathbf{v}}dx\right| \\
&\le \|\tfrac{1}{n^2}\|_\infty\|(\nabla+i\mathbf{k})\times\mathbf{u}\|_{(L_p^2)^3}\|(\nabla+i\mathbf{k})\times\mathbf{v}\|_{(L_p^2)^3} \\
&\le \left(\|\nabla\times\mathbf{u}\|_{(L_p^2)^3}+|\mathbf{k}|\|\mathbf{u}\|_{(L_p^2)^3}\right)\left(\|\nabla\times\mathbf{v}\|_{(L_p^2)^3}+|\mathbf{k}|\|\mathbf{v}\|_{(L_p^2)^3}\right) \quad \text{since } n^2\ge 1 \\
&\le \max\{1,|\mathbf{k}|^2\}\|\mathbf{u}\|_{S_{\mathbf{k}}}\|\mathbf{v}\|_{S_{\mathbf{k}}} \\
&\le 3\pi^2\|\mathbf{u}\|_{S_{\mathbf{k}}}\|\mathbf{v}\|_{S_{\mathbf{k}}}.
\end{aligned}$$

From the definition of $a(\cdot,\cdot)$, it is obvious that $a(\mathbf{u},\mathbf{v})=\overline{a(\mathbf{v},\mathbf{u})}$ for all $\mathbf{u},\mathbf{v}\in F_{\mathbf{k}}$ and so $a(\cdot,\cdot)$ is Hermitian on $F_{\mathbf{k}}$.

Now let us show that $a(\cdot,\cdot)$ satisfies (6.17). For $\mathbf{v}\in F_{\mathbf{k}}$ we get (using the Cauchy-

Schwarz and Arithmetic-Geometric Mean inequalities),

$$\begin{aligned}
a(\mathbf{v},\mathbf{v}) &= \int_\Omega \tfrac{1}{n^2}|(\nabla + i\mathbf{k})\times\mathbf{v}|^2 dx \geq \tfrac{1}{n_{max}^2}\int_\Omega |(\nabla+i\mathbf{k})\times\mathbf{v}|^2 dx \\
&\geq \tfrac{1}{n_{max}^2}\int_\Omega \left(|\nabla\times\mathbf{v}| - |\mathbf{k}||\mathbf{v}|\right)^2 dx \qquad \text{since } |a+b|\geq||a|-|b|| \\
&= \tfrac{1}{n_{max}^2}\int_\Omega |\nabla\times\mathbf{v}|^2 - 2|\mathbf{k}||\nabla\times\mathbf{v}||\mathbf{v}| + |\mathbf{k}|^2|\mathbf{v}|^2 dx \\
&= \tfrac{1}{n_{max}^2}\left(\|\nabla\times\mathbf{v}\|^2_{(L^2_p)^3} + |\mathbf{k}|^2\|\mathbf{v}\|^2_{(L^2_p)^3} - \|\nabla\times\mathbf{v}\|_{(L^2_p)^3}\left(2|\mathbf{k}|\|\mathbf{v}\|_{(L^2_p)^3}\right)\right) \\
&\geq \tfrac{1}{n_{max}^2}\left(\|\nabla\times\mathbf{v}\|^2_{(L^2_p)^3} + |\mathbf{k}|^2\|\mathbf{v}\|^2_{(L^2_p)^3} - \tfrac{1}{2}\|\nabla\times\mathbf{v}\|^2_{(L^2_p)^3} - 2|\mathbf{k}|^2\|\mathbf{v}\|^2_{(L^2_p)^3}\right) \\
&= \tfrac{1}{2n_{max}^2}\|\nabla\times\mathbf{v}\|^2_{(L^2_p)^3} - \tfrac{|\mathbf{k}|^2}{n_{max}^2}\|\mathbf{v}\|^2_{(L^2_p)^3} \\
&\geq \tfrac{1}{2n_{max}^2}\|\nabla\times\mathbf{v}\|^2_{(L^2_p)^3} - \tfrac{3\pi^2}{n_{max}^2}\|\mathbf{v}\|^2_{(L^2_p)^3} \\
&= \tfrac{1}{2n_{max}^2}\|\mathbf{v}\|^2_{S_\mathbf{k}} - \tfrac{6\pi^2+1}{2n_{max}^2}\|\mathbf{v}\|^2_{(L^2_p)^3}.
\end{aligned}$$

Therefore, $a(\cdot,\cdot)$ satisfies (6.17) □

Now we can use Lemmas 6.4 and 6.5 to prove the existence of eigenpairs for Problem 6.3 as well as a regularity result for the eigenfunctions of Problem 6.3.

Theorem 6.6. *Problem 6.3 has real eigenvalues*

$$-\tfrac{6\pi^2+1}{2n_{max}^2} < \lambda_1 \leq \lambda_2 \leq \ldots \nearrow +\infty$$

with corresponding eigenfunctions $\mathbf{u}_1, \mathbf{u}_2, \ldots \in F_\mathbf{k}$ *that satisfy*

$$(\nabla + i\mathbf{k})\times(\tfrac{1}{n^2}(\nabla+i\mathbf{k})\times\mathbf{u}_j) \in F_\mathbf{k} \qquad \textit{for } j = 1,2,\ldots$$

Proof. Define an operator $F : F_\mathbf{k} \to F_\mathbf{k}$ such that

$$a(F\mathbf{u},\mathbf{v}) + (\tfrac{6\pi^2+1}{2n_{max}^2})(F\mathbf{u},\mathbf{v})_{(L^2_p)^3} = b(\mathbf{u},\mathbf{v}) \qquad \forall\mathbf{v}\in F_\mathbf{k}.$$

From Lemma 6.5 and the Lax-Milgram Lemma we know that F is well-defined and $\|F\mathbf{u}\|_{S_\mathbf{k}} \lesssim \|\mathbf{u}\|_{(L^2_p)^3}$. This, together with the fact that $F_\mathbf{k} \subset\subset (L^2_p)^3$ implies that F is compact. We can also show that F is self-adjoint with respect to $a(\cdot,\cdot)+(\frac{6\pi^2+1}{2n_{max}^2})(\cdot,\cdot)_{(L^2_p)^3}$ by using the fact that $a(\cdot,\cdot)$ is Hermitian (see Lemma 6.5). Therefore, by Theorem 3.60, $\sigma(F)$ consists of real eigenvalues, μ_j, of finite multiplicity with the only possible accumulation point at zero, i.e.

$$\mu_1 \geq \mu_2 \geq \ldots > 0.$$

It is easy to show (c.f. Lemma 3.71) that if $(\mu,\mathbf{u})$ is an eigenpair of F then $(\frac{1}{\mu} - \frac{6\pi^2+1}{2n_{max}^2}, u)$

is an eigenpair of Problem 6.3. Therefore, Problem 6.3 has real eigenvalues

$$-\tfrac{6\pi^2+1}{2n_{max}^2} < \lambda_1 \le \lambda_2 \le \ldots \nearrow +\infty$$

where $\lambda_j = \frac{1}{\mu_j} - \frac{6\pi^2+1}{2n_{max}^2}$ for $j \in \mathbb{N}$.

Now let $(\lambda, \mathbf{u})$ be an eigenpair of Problem 6.3. Using the following two properties of functions in $G_{\mathbf{k}}$,

$$\begin{aligned} (\nabla + i\mathbf{k}) \times \mathbf{v} &= \mathbf{0} && \text{for all } \mathbf{v} \in G_{\mathbf{k}} \\ \int_\Omega \mathbf{u} \cdot \mathbf{v} &= 0 && \text{for all } \mathbf{u} \in F_{\mathbf{k}}, \mathbf{v} \in G_{\mathbf{k}} \end{aligned}$$

and Part 4 of Lemma 6.4 we have

$$a(\mathbf{u}, \mathbf{v}) = \lambda b(\mathbf{u}, \mathbf{v}) \qquad \forall \mathbf{v} \in (L_p^2)^3.$$

Therefore,

$$(\nabla + i\mathbf{k}) \times (\tfrac{1}{n^2}(\nabla + i\mathbf{k}) \times \mathbf{u}) = \lambda \mathbf{u} \tag{6.18}$$

in the distributional sense. Since $\mathbf{u} \in F_{\mathbf{k}}$ we get $(\nabla + i\mathbf{k}) \times (\frac{1}{n^2}(\nabla + i\mathbf{k}) \times \mathbf{u}) \in F_{\mathbf{k}}$. □

We would now like to use what we know about Problem 6.3 to try and prove a result about the existence and regularity of eigenpairs of the Full 2D Problem. Our first task is to relate an eigenpair of Problem 6.3 to an eigenpair of (6.4). Unfortunately, the following result is "one-way". It remains an open problem to prove that an eigenpair of (6.4) (in the distributional sense) is an eigenpair of Problem 6.3.

Recall our notation convention, if $\mathbf{v} \in \mathbb{R}^3$ with $\mathbf{v} = (v_1, v_2, v_3)$ then $\mathbf{v}_t := (v_1, v_2, 0)$, $\mathbf{v}_z := (0, 0, v_3)$ and $v_z := v_3$.

Theorem 6.7. *Let $\mathbf{k} \in Q = [-\pi, \pi]^3$ and suppose that $(\lambda, \mathbf{w})$ is an eigenpair of Problem 6.3. Then there exists an $m \in \mathbb{Z}$ such that*

$$\widehat{\mathbf{w}}(x, y; m) = \int_{-1/2}^{1/2} \mathbf{w}(x, y, z)\, \mathrm{e}^{-i2\pi m z}\, dz \neq 0 \tag{6.19}$$

and $(\beta^2, \widehat{\mathbf{w}}_t)$ is an eigenpair of (6.4) with $\boldsymbol{\xi} = \mathbf{k}_t$, $\beta = k_z + 2\pi m$ and $\gamma(\mathbf{x}) = \lambda n^2(\mathbf{x})$.

Proof. Let $\mathbf{k} \in Q$ and suppose $(\lambda, \mathbf{w})$ is an eigenpair of Problem 6.3. Then (as in (6.18)) $(\lambda, \mathbf{w})$ satisfies

$$\begin{aligned} (\nabla + i\mathbf{k}) \times (\tfrac{1}{n^2}(\nabla + i\mathbf{k}) \times \mathbf{w}) &= \lambda \mathbf{w} \\ (\nabla + i\mathbf{k}) \cdot \mathbf{w} &= 0 \end{aligned} \tag{6.20}$$

in $(\mathcal{D}_p'(\mathbb{R}^3))^3$, i.e. in the periodic distributional sense. For the rest of this proof we simplify our notation and just write $\mathcal{D}_p'(\mathbb{R}^d)$ to mean $(\mathcal{D}_p'(\mathbb{R}^d))^3$. Since $\mathbf{w}$ is a periodic

distribution with respect to z we can expand it in terms of its Fourier Series to get

$$\mathbf{w}(x,y,z) = \sum_{r\in\mathbb{Z}} \widehat{\mathbf{w}}(x,y;r)\,\mathrm{e}^{i2\pi rz} \qquad \text{in } \mathcal{D}'_p(\mathbb{R}^3)$$

where

$$\widehat{\mathbf{w}}(x,y;r) = \int_{-1/2}^{1/2} \mathbf{w}(x,y,z)\,\mathrm{e}^{-i2\pi rz}\,dz.$$

Substituting this expansion of $\mathbf{w}$ into (6.20) we get

$$\sum_{r\in\mathbb{Z}} (\nabla + i\mathbf{k}) \times (\tfrac{1}{n^2}(\nabla + i\mathbf{k}) \times (\widehat{\mathbf{w}}(x,y;r)\,\mathrm{e}^{i2\pi rz})) = \lambda \sum_{r\in\mathbb{Z}} \widehat{\mathbf{w}}(x,y;r)\,\mathrm{e}^{i2\pi rz}$$
$$\sum_{r\in\mathbb{Z}} (\nabla + i\mathbf{k}) \cdot (\widehat{\mathbf{w}}(x,y;r)\,\mathrm{e}^{i2\pi rz}) = 0 \qquad \text{in } \mathcal{D}'_p(\mathbb{R}^3).$$

Using the product rule we then get

$$\begin{aligned}\sum_{r\in\mathbb{Z}} \big[(\nabla_t + i\mathbf{k} + i2\pi r\hat{\mathbf{z}}) \times (\tfrac{1}{n^2}(\nabla_t + i\mathbf{k} + i2\pi r\hat{\mathbf{z}}) \times \widehat{\mathbf{w}}(x,y;r))\big]\,\mathrm{e}^{i2\pi rz} \\ = \lambda \sum_{r\in\mathbb{Z}} \widehat{\mathbf{w}}(x,y;r)\,\mathrm{e}^{i2\pi rz} \\ \sum_{r\in\mathbb{Z}} \big[(\nabla_t + i\mathbf{k} + i2\pi r\hat{\mathbf{z}}) \cdot \widehat{\mathbf{w}}(x,y;r)\big]\,\mathrm{e}^{i2\pi rz} = 0 \qquad \text{in } \mathcal{D}'_p(\mathbb{R}^3).\end{aligned} \tag{6.21}$$

Since $\mathbf{w} \neq 0$ there exists an $m \in \mathbb{Z}$ such that $\widehat{\mathbf{w}}(x,y;m) \neq 0$. By matching the Fourier coefficients (for $r = m$) in (6.21) we obtain

$$(\nabla_t + i\mathbf{k} + i2\pi m\hat{\mathbf{z}}) \times (\tfrac{1}{n^2}(\nabla_t + i\mathbf{k} + i2\pi m\hat{\mathbf{z}}) \times \widehat{\mathbf{w}}(x,y;m)) = \lambda \widehat{\mathbf{w}}(x,y;m)$$
$$(\nabla_t + i\mathbf{k} + i2\pi m\hat{\mathbf{z}}) \cdot \widehat{\mathbf{w}}(x,y;m) = 0 \qquad \text{in } (\mathcal{D}'_p(\mathbb{R}^2))^3.$$

Now set $\boldsymbol{\xi} = \mathbf{k}_t$ and $\beta = k_z + 2\pi m$ (and let $\widehat{\mathbf{w}} = \widehat{\mathbf{w}}(x,y;m)$) to get

$$(\nabla_t + i\boldsymbol{\xi} + i\beta\hat{\mathbf{z}}) \times (\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi} + i\beta\hat{\mathbf{z}}) \times \widehat{\mathbf{w}}) = \lambda \widehat{\mathbf{w}}$$
$$(\nabla_t + i\boldsymbol{\xi} + i\beta\hat{\mathbf{z}}) \cdot \widehat{\mathbf{w}} = 0 \qquad \text{in } (\mathcal{D}'_p(\mathbb{R}^2))^3.$$

Now split the first equation into transverse and z components to get (after cancelling terms that are zero)

$$\begin{aligned}(\nabla_t + i\boldsymbol{\xi}) \times (\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_t) + i\beta\hat{\mathbf{z}} \times (\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_z) \\ + i\beta\hat{\mathbf{z}} \times (\tfrac{1}{n^2} i\beta\hat{\mathbf{z}} \times \widehat{\mathbf{w}}_t) = \lambda \widehat{\mathbf{w}}_t \\ (\nabla_t + i\boldsymbol{\xi}) \times (\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_z) + (\nabla_t + i\boldsymbol{\xi}) \times (\tfrac{1}{n^2} i\beta\hat{\mathbf{z}} \times \widehat{\mathbf{w}}_t) = \lambda \widehat{\mathbf{w}}_z \\ (\nabla_t + i\boldsymbol{\xi}) \cdot \widehat{\mathbf{w}}_t + i\beta\widehat{w}_z = 0 \qquad \text{in } (\mathcal{D}'_p(\mathbb{R}^2))^3.\end{aligned} \tag{6.22}$$

Now use the following identities

$$i\beta\hat{\mathbf{z}} \times (\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_z) = \tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi})(i\beta\widehat{w}_z)$$
$$i\beta\hat{\mathbf{z}} \times (\tfrac{1}{n^2}i\beta\hat{\mathbf{z}} \times \widehat{\mathbf{w}}_t) = \tfrac{1}{n^2}\beta^2\widehat{\mathbf{w}}_t$$

to simplify (6.22) to get

$$\begin{aligned}(\nabla_t + i\boldsymbol{\xi}) \times (\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_t) + \tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi})(i\beta\widehat{w}_z) + \tfrac{1}{n^2}\beta^2\widehat{\mathbf{w}}_t = \lambda\widehat{\mathbf{w}}_t \qquad (6.23)\\ (\nabla_t + i\boldsymbol{\xi}) \times (\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_z) + (\nabla_t + i\boldsymbol{\xi}) \times (\tfrac{1}{n^2}i\beta\hat{\mathbf{z}} \times \widehat{\mathbf{w}}_t) = \lambda\widehat{\mathbf{w}}_z\\ (\nabla_t + i\boldsymbol{\xi}) \cdot \widehat{\mathbf{w}}_t + i\beta\widehat{\mathbf{w}}_z = 0 \qquad \text{in } (\mathcal{D}'_p(\mathbb{R}^2))^3\end{aligned}$$

Now substitute $i\beta\widehat{\mathbf{w}}_z = -(\nabla_t + i\boldsymbol{\xi}) \cdot \widehat{\mathbf{w}}_t$ into (6.23) and expand the first term using the product rule to get

$$\begin{aligned}\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi}) \times ((\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_t) + \nabla_t(\tfrac{1}{n^2}) \times ((\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_t)\\ -\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi})((\nabla_t + i\boldsymbol{\xi}) \cdot \widehat{\mathbf{w}}_t) + \tfrac{1}{n^2}\beta^2\widehat{\mathbf{w}}_t = \lambda\widehat{\mathbf{w}}_t \qquad \text{in } (\mathcal{D}'_p(\mathbb{R}^2))^3.\end{aligned} \qquad (6.24)$$

Now use the identity

$$(\nabla_t + i\boldsymbol{\xi}) \times ((\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_t) - (\nabla_t + i\boldsymbol{\xi})((\nabla_t + i\boldsymbol{\xi}) \cdot \widehat{\mathbf{w}}_t) = -(\nabla + i\boldsymbol{\xi})^2\widehat{\mathbf{w}}_t$$

to simplify (6.24) to get

$$-\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi})^2\widehat{\mathbf{w}}_t + \nabla_t(\tfrac{1}{n^2}) \times ((\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_t) + \tfrac{1}{n^2}\beta^2\widehat{\mathbf{w}}_t = \lambda\widehat{\mathbf{w}}_t \qquad \text{in } (\mathcal{D}'_p(\mathbb{R}^2))^3.$$

Multiplying by $-n^2$ and rearranging terms we get

$$(\nabla_t + i\boldsymbol{\xi})^2\widehat{\mathbf{w}}_t + \lambda n^2\widehat{\mathbf{w}}_t - (n^2\nabla_t(\tfrac{1}{n^2})) \times ((\nabla_t + i\boldsymbol{\xi}) \times \widehat{\mathbf{w}}_t) + = \beta^2\widehat{\mathbf{w}}_t \qquad \text{in } (\mathcal{D}'_p(\mathbb{R}^2))^3.$$

With $-n^2\nabla_t(\frac{1}{n^2}) = \nabla_t(\log n^2)$ we have that $(\beta^2, \widehat{\mathbf{w}}_t)$ is an eigenpair of (6.4) (in the distributional sense) with $\boldsymbol{\xi} = \mathbf{k}_t$, $\beta = k_z + 2\pi m$ and $\gamma(\mathbf{x}) = \lambda n^2(\mathbf{x})$. □

If we consider the converse argument then it is possible to show that if $(\beta^2, \mathbf{u})$ is an eigenpair of (6.4) for some $\boldsymbol{\xi} \in B$ and $\beta^2 \geq 0$ (in the distributional sense) where $\gamma = \lambda n^2$ then there exists an $m \in \mathbb{Z}$ such that $k_z = \beta - 2\pi m \in [-\pi, \pi]$ and $(\lambda, \mathbf{w})$ is an eigenpair of (6.20) (also in the distributional sense) where $\mathbf{k} = (\xi_1, \xi_2, k_z)$ and $\mathbf{w}(x, y, z) = \widehat{\mathbf{w}}(x, y)\,\mathrm{e}^{i2\pi mz}$, with $\widehat{\mathbf{w}} := (u_1, u_2, \frac{i}{\beta}(\nabla_t + i\boldsymbol{\xi}) \cdot \mathbf{u})$. Unfortunately, the converse arguement then fails because a distributional solution to (6.20) is not necessarily a solution to Problem 6.3 since it lacks regularity.

Nevertheless, using Theorem 6.6 and Theorem 6.7 together ensures the existence of eigenpairs of (6.4) (in the distributional sense) and that these eigenpairs correspond

to eigenpairs of Problem 6.3. For the rest of this chapter we restrict our attention to eigenpairs of (6.4) that are also eigenpairs of Problem 6.3.

Lemma 6.8. *Let $\boldsymbol{\xi} \in B$ and let $(\beta^2, \mathbf{u})$ be an eigenpair of (6.4) with $\gamma = \lambda n^2$ such that $(\lambda, \mathbf{w})$ is a corresponding eigenpair of Problem 6.3 (i.e there exists an eigenpair of Problem 6.3 such that Theorem 6.7 implies that $(\beta^2, \mathbf{u})$ is an eigenpair of (6.4)). Then $\mathbf{u}(x, y, z) = \widetilde{\mathbf{w}}_t(x, y, z)\, \mathrm{e}^{-i2\pi mz}$ where $\widetilde{\mathbf{w}}$ is an eigenfunction of Problem 6.3 (possibly different from $\mathbf{w}$) and $m \in \mathbb{Z}$ is defined in Theorem 6.7. Moreover, $\mathbf{u} = (u_1, u_2, 0) \in (L_p^2)^3$ and $(\nabla_t + i\boldsymbol{\xi}) \times \mathbf{u} \in (L_p^2)^3$.*

Proof. Since $(\beta^2, \mathbf{u})$ corresponds to an eigenpair of Problem 6.3 there exists an eigenpair of Problem 6.3 $(\lambda, \mathbf{w})$ for some $m \in \mathbb{Z}$ such that $\mathbf{k} = (\xi_1, \xi_2, \beta - 2\pi m)$, and $\mathbf{u}(x, y) = \widehat{\mathbf{w}}_t(x, y; m)$ where $\widehat{\mathbf{w}}$ is defined in (6.24).

Using similar steps to the proof of Theorem 6.7, but in reverse, we can show that $(\lambda, \widetilde{\mathbf{w}})$ where $\widetilde{\mathbf{w}}(x, y, z) = \widehat{\mathbf{w}}(x, y; m)\, \mathrm{e}^{i2\pi mz}$ is an eigenpair (in the distributional sense) of (6.20). We can also show that $\widetilde{\mathbf{w}}$ possesses sufficient regularity so that $(\lambda, \widetilde{\mathbf{w}})$ is an eigenfunction of Problem 6.3. For this we need to show that $\widetilde{\mathbf{w}} \in F_{\mathbf{k}}$, i.e. we need to show that $\widetilde{\mathbf{w}} \in (L_p^2)^3$, $\nabla \times \widetilde{\mathbf{w}} \in (L_p^2)^3$ and $(\nabla + \mathbf{k}) \cdot \widetilde{\mathbf{w}} = 0$ (this follows directly from (6.20) using a density argument). By writing $\widetilde{\mathbf{w}}$ as

$$\widetilde{\mathbf{w}}(\mathbf{x}) = \sum_{\substack{\mathbf{g} \in \mathbb{Z}^3 \\ g_3 = m}} [\mathbf{w}]_{\mathbf{g}}\, \mathrm{e}^{i2\pi \mathbf{g} \cdot \mathbf{x}}$$

it then follows directly from the definition of the H_p^s norm and the linearity of $\nabla\times$ that $\|\widetilde{\mathbf{w}}\|_{(H_p^s)^3} \leq \|\mathbf{w}\|_{(H_p^s)^3}$ and $\|\nabla \times \widetilde{\mathbf{w}}\|_{(H_p^s)^3} \leq \|\nabla \times \mathbf{w}\|_{(H_p^s)^3}$ for all $s \in \mathbb{R}$. Thus, with $s = 0$ we have shown that $\widetilde{\mathbf{w}} \in F_{\mathbf{k}}$ and it then follows from (6.20) by a density argument that $(\lambda, \widetilde{\mathbf{w}})$ is an eigenfunction of Problem 6.3.

By the correspondence between $\mathbf{w}$ and $\mathbf{u}$ defined in Theorem 6.7 (and a slight abuse of notation)

$$\mathbf{u}(x, y) = \widehat{\mathbf{w}}_t(x, y; m) = \widetilde{\mathbf{w}}_t(x, y, z)\, \mathrm{e}^{-i2\pi mz}$$

Since $\mathbf{u}(x, y, z) = \widetilde{\mathbf{w}}_t(x, y, z)\, \mathrm{e}^{-i2\pi mz}$ and $\widetilde{\mathbf{w}} \in (L_p^2)^3$ it follows that $\mathbf{u} \in (L_p^2)^3$.

Moreover, since $\widetilde{\mathbf{w}} \in F_{\mathbf{k}}$ we have

$$(L_p^2)^3 \ni (\nabla + i\mathbf{k}) \times \widetilde{\mathbf{w}} = \begin{bmatrix} \frac{\partial \widetilde{w}_3}{\partial y} - \frac{\partial \widetilde{w}_2}{\partial z} \\ \frac{\partial \widetilde{w}_1}{\partial z} - \frac{\partial \widetilde{w}_3}{\partial x} \\ \frac{\partial \widetilde{w}_2}{\partial x} - \frac{\partial \widetilde{w}_1}{\partial y} \end{bmatrix} + i\mathbf{k} \times \widetilde{\mathbf{w}} = \begin{bmatrix} \frac{\partial \widetilde{w}_3}{\partial y} - i2\pi m \widetilde{w}_2 \\ i2\pi m \widetilde{w}_1 - \frac{\partial \widetilde{w}_3}{\partial x} \\ \frac{\partial \widetilde{w}_2}{\partial x} - \frac{\partial \widetilde{w}_1}{\partial y} \end{bmatrix} + i \begin{bmatrix} \xi_2 \widetilde{w}_3 - k_z \widetilde{w}_2 \\ k_z \widetilde{w}_1 - \xi_1 \widetilde{w}_3 \\ \xi_1 \widetilde{w}_2 - \xi_2 \widetilde{w}_1 \end{bmatrix}$$

which implies that $\frac{\partial \widetilde{w}_3}{\partial x} \in L_p^2$ and $\frac{\partial \widetilde{w}_3}{\partial y} \in L_p^2$. We also have $\frac{\partial \widetilde{w}_3}{\partial z} = i2\pi m \widetilde{w}_3 \in L_p^2$ and so it follows that $\widetilde{w}_3 \in H_p^1$. Moreover, using the above expressions we can show that

$$\|\widetilde{w}_3\|_{H_p^1} \lesssim \|\widetilde{\mathbf{w}}\|_{S_{\mathbf{k}}}. \tag{6.25}$$

Therefore,

$$\begin{aligned}
\|(\nabla_t + i\boldsymbol{\xi}) \times \mathbf{u}\|_{(L^2_p)^3} &= \| \mathrm{e}^{-i2\pi mz}(\nabla_t + i\mathbf{k}_t) \times \widetilde{\mathbf{w}}_t\|_{(L^2_p)^3} \\
&= \|(\nabla_t + i\mathbf{k}_t) \times \widetilde{\mathbf{w}}_t\|_{(L^2_p)^3} \\
&= \|(\nabla + i\mathbf{k}) \times \widetilde{\mathbf{w}} - (\nabla_t + i\mathbf{k}_t) \times \widetilde{\mathbf{w}}_z - (\nabla_z + i\mathbf{k}_z) \times \widetilde{\mathbf{w}}_t\|_{(L^2_p)^3} \\
&\leq \|(\nabla + i\mathbf{k}) \times \widetilde{\mathbf{w}}\|_{(L^2_p)^3} + \|(\nabla_t + i\mathbf{k}_t) \times \widetilde{\mathbf{w}}_z\|_{(L^2_p)^3} \\
&\qquad + \|(\nabla_z + i\mathbf{k}_z) \times \widetilde{\mathbf{w}}_t\|_{(L^2_p)^3} \\
&\lesssim \|\widetilde{\mathbf{w}}\|_{S_\mathbf{k}} + \|\widetilde{w}_3\|_{H^1_p} + \|\widetilde{\mathbf{w}}_t\|_{(L^2_p)^3} \\
&\lesssim \|\widetilde{\mathbf{w}}\|_{S_\mathbf{k}}
\end{aligned}$$

and $(\nabla_t + i\boldsymbol{\xi}) \times \mathbf{u} \in (L^2_p)^3$. □

We now prove another result about the regularity of eigenfunctions of (6.4) (that correspond to eigenfunctions of Problem 6.3).

Theorem 6.9. *Let $\boldsymbol{\xi} \in B$ and let $(\beta^2, \mathbf{u})$ be an eigenpair of (6.4) with $\gamma = \lambda n^2$ such that $(\lambda, \mathbf{w})$ is a corresponding eigenpair of Problem 6.3 (i.e there exists an eigenpair of Problem 6.3 such that Theorem 6.7 implies that $(\beta^2, \mathbf{u})$ is an eigenpair of (6.4)). Then there exists $s \in \mathbb{R}$ with $s \geq 0$ such that $\mathbf{u} \in (H^{1+s}_p)^3$ (recall that $u_3 = 0$).*

Proof. Rewrite (6.4) as a 2D elliptic boundary value problem: Find $\mathbf{u} = (u_1, u_2, 0) \in (H^1_p)^3$ such that

$$L\mathbf{u} = \mathbf{f} \qquad \text{on } \mathbb{R}^2 \tag{6.26}$$

where

$$\begin{aligned}
L &:= -(\nabla + i\boldsymbol{\xi})^2 = -\nabla^2 - 2i\boldsymbol{\xi} \cdot \nabla + |\boldsymbol{\xi}|^2 \\
\mathbf{f} &:= -\beta^2\mathbf{u} - \gamma\mathbf{u} - (\nabla_t\eta) \times ((\nabla_t + i\boldsymbol{\xi}) \times \mathbf{u}).
\end{aligned}$$

Notice that L is elliptic (definition in Section 3.5.5) and has constant coefficients. Also notice that we can separate (6.26) into the components $Lu_1 = f_1$ and $Lu_2 = f_2$ ($Lu_3 = f_3$ is meaningless because $u_3 = f_3 = 0$).

If we can show that $\mathbf{f} \in (H^{-1+s}_p)^3$ for some $s \geq 0$ then we can prove the result using Theorem 3.2 on page 125 of [52] which says: For $r \in \mathbb{Z}$, if L is 2nd-order and elliptic with infinitely differentiable coefficients and $Lu \in H^{r-2}(\widetilde{\Omega})$, then $u \in H^r_{loc}(\widetilde{\Omega})$. Note Remark 3.2 on page 127 of [52] which says that Theorem 3.2 applies for $r \in \mathbb{R}$.

We can apply this theorem to both $Lu_1 = f_1$ and $Lu_2 = f_2$ by choosing $\widetilde{\Omega}$ so that $\widetilde{\Omega}$ is bounded and $\Omega \subset\subset \widetilde{\Omega}$.

It remains to show that $\mathbf{f} \in (H^{-1+s}_p)^3$ for some $s \geq 0$. Since $\mathbf{u} \in (L^2_p)^3$ (Lemma 6.8), we also have

$$-\beta^2\mathbf{u} - \gamma\mathbf{u} \in (L^2_p)^3. \tag{6.27}$$

Now let us consider the third term in $\mathbf{f}$.

$$\begin{aligned}
&(\nabla_t\eta)\times((\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}) = \\
&\quad=\left(\tfrac{1}{n^2}\nabla_t n^2\right)\times((\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}) \\
&\quad=\left(\nabla_t n^2\right)\times\left(\tfrac{1}{n^2}(\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}\right) \\
&\quad=\nabla_t\times\left(n^2\left(\tfrac{1}{n^2}(\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}\right)\right)-n^2\nabla_t\times\left(\tfrac{1}{n^2}(\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}\right) \\
&\quad=\underbrace{\nabla_t\times((\nabla_t+i\boldsymbol{\xi})\times\mathbf{u})}_{I_1}-\underbrace{n^2\nabla_t\times\left(\tfrac{1}{n^2}(\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}\right)}_{I_2}.
\end{aligned} \tag{6.28}$$

We will now show that $I_1\in(H_p(curl))^*$ (the dual of $H_p(curl)$) and $I_2\in(L^2_p)^3$.

Let $\mathbf{v}\in H_p(curl)$. Then (with $\boldsymbol{\nu}$ denoting the outward pointing normal on $\partial\Omega$),

$$\begin{aligned}
\int_\Omega I_1\cdot\mathbf{v}\,d\mathbf{x} &= \int_\Omega(\nabla_t\times((\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}))\cdot\mathbf{v}\,d\mathbf{x} \\
&= \int_\Omega(\nabla\times((\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}))\cdot\mathbf{v}\,d\mathbf{x} \qquad \text{since } \mathbf{u}=\mathbf{u}(x,y) \\
&= \int_\Omega(\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}\cdot\nabla\times\mathbf{v}\,d\mathbf{x}+\int_{\partial\Omega}\boldsymbol{\nu}\times((\nabla_t+i\boldsymbol{\xi})\times\mathbf{u})\cdot\mathbf{v}\,d\mathbf{x} \\
&= \int_\Omega(\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}\cdot\nabla\times\mathbf{v}\,d\mathbf{x} \qquad \text{since } \mathbf{u},\mathbf{v} \text{ periodic} \\
&\le \|(\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}\|_{(L^2_p)^3}\|\nabla\times\mathbf{v}\|_{(L^2_p)^3} \qquad \text{by Cauchy-Schwarz} \\
&\le \|(\nabla_t+i\boldsymbol{\xi})\times\mathbf{u}\|_{(L^2_p)^3}\|\mathbf{v}\|_{H_p(curl)}
\end{aligned}$$

Therefore, it follows from Lemma 6.8 that $I_1\in(H_p(curl))^*$.

Now consider I_2. It follows from Lemma 6.8 that $\mathbf{u}(x,y,z)=\widetilde{\mathbf{w}}_t(x,y,z)\,\mathrm{e}^{-i2\pi mz}$ where $\widetilde{\mathbf{w}}(x,y,z):=\widehat{\mathbf{w}}(x,y;m)\,\mathrm{e}^{i2\pi mz}$ ($\widehat{\mathbf{w}}$ defined in (6.24)) is an eigenfunction of Problem 6.3 and $m\in\mathbb{Z}$.

In the following argument let us define functions $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$ and $\mathbf{f}^{(3)}$ by

$$\begin{aligned}
\mathbf{f}^{(1)} &:= \tfrac{1}{n^2}(\nabla+i\mathbf{k})\times\widetilde{\mathbf{w}} \\
\mathbf{f}^{(2)} &:= (\nabla+i\mathbf{k})\times\mathbf{f}^{(1)} \\
\mathbf{f}^{(3)} &:= \nabla\times\mathbf{f}^{(1)}.
\end{aligned}$$

Since $\widetilde{\mathbf{w}}\in F_{\mathbf{k}}$, it follows that $\mathbf{f}^{(1)}\in(L^2_P)^3$. Theorem 6.6 implies that $\mathbf{f}^{(2)}\in F_{\mathbf{k}}$. It then follows that $\mathbf{f}^{(3)}=\mathbf{f}^{(2)}-i\mathbf{k}\times\mathbf{f}^{(1)}\in(L^2_p)^3$.

Using the relationship between $\mathbf{u}$ and $\mathbf{w}_t$ and our definitions of $\mathbf{f}^{(i)}$, we get

$$\begin{aligned}
&\|\nabla_t \times (\tfrac{1}{n^2}(\nabla_t + i\boldsymbol{\xi}) \times \mathbf{u})\|_{(L_p^2)^3} = \\
&\quad = \|\,\mathrm{e}^{-i2\pi mz}\,\nabla_t \times (\tfrac{1}{n^2}(\nabla_t + i\mathbf{k}_t) \times \widetilde{\mathbf{w}}_t)\|_{(L_p^2)^3} \qquad \text{since } \mathbf{u} = \widetilde{\mathbf{w}}_t\,\mathrm{e}^{i2\pi mz} \\
&\quad = \|\nabla_t \times (\tfrac{1}{n^2}(\nabla_t + i\mathbf{k}_t) \times \widetilde{\mathbf{w}}_t)\|_{(L_p^2)^3} \\
&\quad = \|\mathbf{f}_t^{(3)} - \nabla_z \times (\tfrac{1}{n^2}(\nabla_t + i\mathbf{k}_t) \times \widetilde{\mathbf{w}}_z)\|_{(L_p^2)^3} \qquad \text{by expanding } \mathbf{f}^{(3)}\text{, other terms 0} \\
&\quad \leq \|\mathbf{f}^{(3)}\|_{(L_p^2)^3} + \|\nabla_z \times (\tfrac{1}{n^2}(\nabla_t + i\mathbf{k}_t) \times \widetilde{\mathbf{w}}_z)\|_{(L_p^2)^3} \\
&\quad \lesssim \|\mathbf{f}^{(3)}\|_{(L_p^2)^3} + \|\widetilde{w}_3\|_{H_p^1} \qquad \text{since } \widetilde{w}_3 = \widehat{w}_3(x,y)\,\mathrm{e}^{i2\pi mz} \\
&\quad \lesssim \|\mathbf{f}^{(3)}\|_{(L_p^2)^3} + \|\mathbf{w}\|_{S_{\mathbf{k}}} \qquad \text{by (6.25)} \\
&\quad < \infty \qquad \text{since } \mathbf{f}^{(3)} \in (L_p^2)^3 \text{ (Theorem 6.6).}
\end{aligned}$$

Therefore, $I_2 \in (L_p^2)^3$.

It now follows from (6.27), (6.28), $I_1 \in (H_p(curl))^*$ and $I_2 \in (L_p^2)^3$ that $\mathbf{f} \in (H_p(curl))^*$.

Finally, Lemma 6.4 implies that

$$(H_p(curl))^* \subsetneq ((H_p^1)^3)^* = (H_p^{-1})^3.$$

Therefore, $\mathbf{f} \in (H_p^{1+s})^3$ for some $s \geq 0$. □

In the preceding theorem we would really like to get $u \in (H_p^{1+s})^3$ for some $s > 0$. To get this result we require that

$$H_p(curl)^* \subset (H_p^{-1+\epsilon}) \tag{6.29}$$

for some $\epsilon > 0$. Unfortunately, we do not know of a proof of this result in the literature.

If such a result existed then we could use the following corollary to guarantee that the approximation error for eigenfunctions of (6.4) that correspond to eigenfunctions of Problem 6.3, approximated with functions in $\mathcal{S}_G$ must converge to zero.

Corollary 6.10. *Let* $\mathbf{u}$ *be an eigenfunction of* (6.4) *(that corresponds to an eigenfunction of Problem 6.3 in the sense of Theorem 6.7) and* $G \in \mathbb{N}$*. Then there exists an* $0 \leq s \leq 1/2$ *such that*

$$\inf_{\boldsymbol{\chi} \in (\mathcal{S}_G)^3} \|\mathbf{u} - \boldsymbol{\chi}\|_{(H_p^1)^3} \lesssim G^{-s}.$$

Proof. Choose $\chi = \mathrm{P}_G^{(\mathcal{S})}\,\mathbf{u}$ and use Theorem 3.30 and Theorem 6.9. □

Another result that might be possible to prove is that if $\mathbf{u}$ is an eigenfunction of (6.4) (that corresponds to an eigenfunction of Problem 6.3) then $\mathbf{u} \notin (H_p^{3/2})^3$ but this requires further investigation.

Computing Reference Solutions to Model Problems 3 and 4	
G	$2^9 - 1$
$N = \dim \mathrm{A}$	$\approx 1.5 \times 10^6$
$(N_f)^2$ (FFT size)	2^{24}
Total Memory (Mb)	≈ 1100
CPU time (seconds)	$\mathcal{O}(10^3)$

Table 6.1: The details of computing reference solutions for Model Problems 3 and 4.

Unfortunately, for the reasons given at the beginning of the section, we have not been able to prove the stability of the plane wave expansion method applied to (6.4), i.e. we have not been able to bound the eigenvalue and eigenfunction errors in terms of the approximation error. However, if we assume that this property is true *and* if (6.29) is true then we could show, via a solution operator argument using Theorem 3.68, that the eigenfunction errors are $\mathcal{O}(G^{-s})$ for some $s > 0$. For the eigenvalue errors, we could also use solution operators and the theory from Theorem 3.68 to bound the errors in terms of the approximation error. However, Problem 6.2 is not symmetric so we could not derive a bound for the eigenvalue errors that is smaller than $\mathcal{O}(G^{-s})$.

6.4 Examples

In this section we compute approximations to the Full 2D Problem using the plane wave expansion method by solving (6.11) as an approximation to (6.4). We observe that the eigenvalue and eigenfunction errors decay at rates that are consistent with the regularity results that we proved in the previous section, and the results suggest that ϵ in Theorem 6.9 and Corollary 6.10 can be chosen arbitrarily small.

We do computations for the PCF structures of Model Problems 3 and 4 that we defined in Subsection 4.1.7 for the Scalar 2D Problem. In particular, $n(x, y)$ is piecewise constant with $n(x, y) = 1$ in *air* regions and $n(x, y) = 1.4$ in *glass* regions. Figure 4-2 represents the period cell of $n(x, y)$ for the different model problems. As in previous chapters $\lambda_0 = 0.5$.

To examine the convergence properties of the plane wave expansion method for these two model problems we have solved (6.11) for varying G and we have calculated the errors of the method by comparing the eigenvalues and eigenfunctions against a reference solution. For both model problems the reference solution is the solution to (6.11) with $G = 2^9 - 1$ and we have calculated the H^1_p norm of the error of normalised eigenfunctions and the relative error of eigenvalues. Table 6.1 contains some details from the computation of the reference solutions.

In Figures 6-1 and 6-2 we see that the eigenfunctions converge at least with $\mathcal{O}(G^{-1/2})$ and that the eigenvalues converge with $\mathcal{O}(G^{-1})$. The fact that we observe faster con-

vergence for Model Problem 4 than we do for Model Problem 3 (for the eigenfunctions) is surprising because Model Problem 4 is a more complicated problem. One possible reason for this is that for Model Problem 4 we have not yet entered a truly asymptotic regime for the size of G that we have chosen. Unfortunately, we have reached the limits of how large we can practicably choose G for computations so we were not able to investigate this further.

The observed decay rate for the eigenfunction errors, $\mathcal{O}(G^{-1/2})$, is the same rate that the approximation error decays at in Corollary 6.10 when we choose $s = 1/2$. This suggests that not only is the plane wave expansion method stable for eigenfunctions (i.e. we can bound the error in terms of the approximation error for plane waves), but the regularity result in Theorem 6.9 should be true for *all* $0 \leq s \leq 1/2$.

The observed decay rate for the eigenvalue errors, $\mathcal{O}(G^{-1})$, is twice as fast as the eigenfunction error, and confirms the conclusion that the plane wave expansion method is stable. Moreover, it also suggests that there is a certain degree of symmetry to the plane wave expansion method for this problem (even though (6.11) is a non-symmetric eigenproblem) since the eigenvalue errors decay at twice the rate of the eigenfunction errors. Recall that in Chapter 4 we saw this behaviour for cases when the continuous and discrete problems were symmetric.

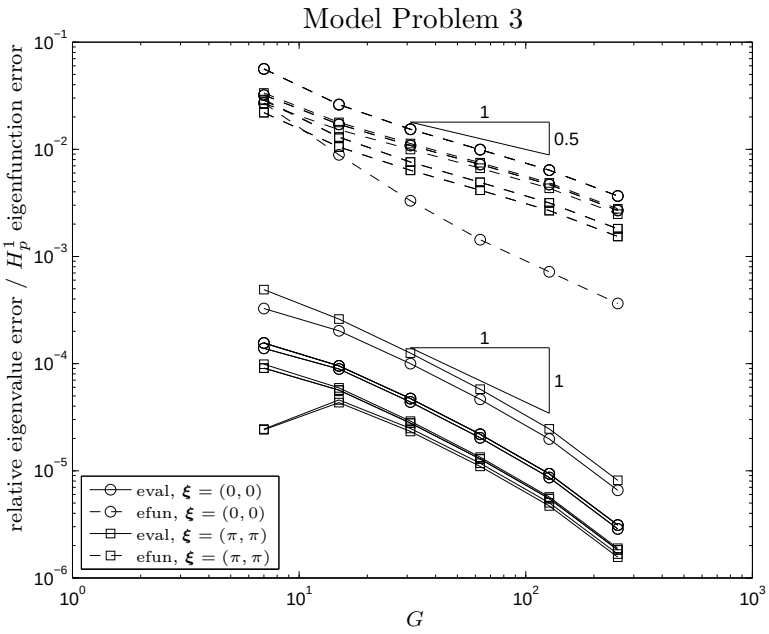

Figure 6-1: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. G for the first 6 eigenpairs of Model Problem 3 (solved for both $\boldsymbol{\xi} = (0,0)$ and $\boldsymbol{\xi} = (\pi,\pi)$).

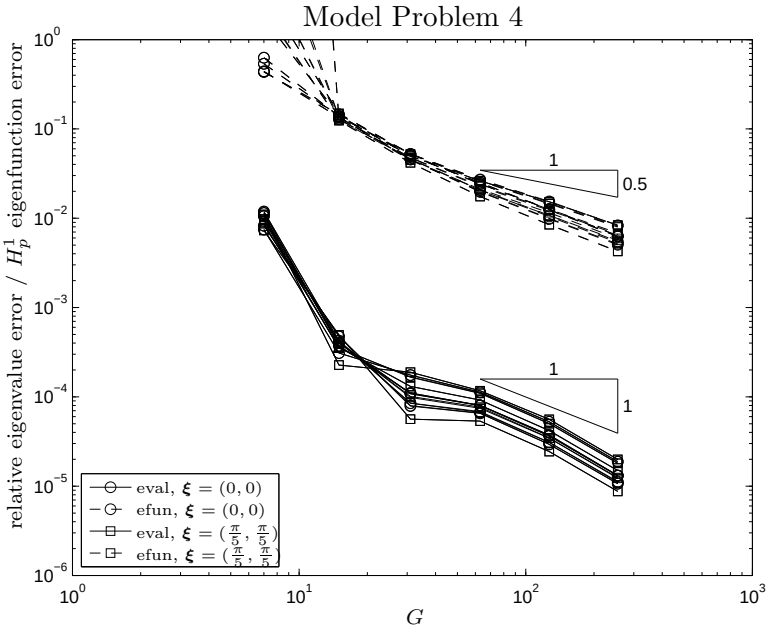

Figure 6-2: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. G for the 21st-30th eigenpairs of Model Problem 4 (solved for both $\boldsymbol{\xi} = (0,0)$ and $\boldsymbol{\xi} = (\frac{\pi}{5},\frac{\pi}{5})$).

6.5 Other Examples: Smoothing and Sampling

In our final section of this chapter we briefly consider smoothing and sampling with the plane wave expansion method for the Full 2D Problem. We would like to know whether or not the conclusions we made about these methods for the other problems extend to the Full 2D Problem. In particular, we would like to know if smoothing is of any benefit to the plane wave expansion method and how fine we should choose our sampling grid to recover the accuracy of the standard plane wave expansion method (that is implemented with exact Fourier coefficients).

We have already applied smoothing and sampling in Sections 4.3, 4.4 and 5.5 for the other problems and the methods are no different here. To implement the smoothing method we solve (6.11) with $[\gamma]_{\mathbf{g}}$ and $[\eta]_{\mathbf{g}}$ in the definition of A replaced with $\mathrm{e}^{i2\pi^2|\mathbf{g}|^2\Delta^2}[\gamma]_{\mathbf{g}}$ and $\mathrm{e}^{i2\pi^2|\mathbf{g}|^2\Delta^2}[\eta]_{\mathbf{g}}$ respectively, where Δ is the parameter that determines the amount of smoothing.

To implement the sampling method we solve (6.11) with $[\gamma]_{\mathbf{g}}$ and $[\eta]_{\mathbf{g}}$ in the definition of A replaced with $[\mathrm{Q}_M\,\gamma]_{\mathbf{g}}$ and $[\mathrm{Q}_M\,\eta]_{\mathbf{g}}$ respectively, where Q_M is the Interpolation Projection defined in Subsection 3.2.5 and $M \in \mathbb{N}$ is the inverse of the grid spacing for the sampling grid.

In all of our plots in this section we have calculated the relative eigenvalue error and H^1_p norm of the error of normalised eigenfunctions, and in all of the plots the reference solution is the solution to (6.11) with $G = 2^9 - 1$, no smoothing and exact Fourier coefficients. See Table 6.1 for some of the details for computing these reference solutions. When we apply the sampling method there will be an additional memory requirement of an $M \times M$ complex double matrix. The largest M that we compute with is $M = 2^{13}$ and this corresponds to an additional 1Gb of memory.

First, let us discuss the smoothing method results. In Figures 6-3 and 6-4 we have plotted the errors for fixed $G = 2^8 - 1$ and varying amounts of smoothing, i.e. varying Δ. In both plots we clearly see that the eigenfunctions decay with $\mathcal{O}(\Delta)$ while the eigenvalues decay with $\mathcal{O}(\Delta^2)$. These results suggest that, to ensure that the smoothing error is less than or equal to the plane wave expansion method error ($\mathcal{O}(G^{-1/2})$ for eigenfunctions and $\mathcal{O}(G^{-1})$ for eigenvalues) in the asymptotic limit, we should choose $\Delta \lesssim G^{-1/2}$.

In Figures 6-5 and 6-6 we have experimented with choosing $\Delta = G^r$ for different constants r. In Figure 6-5 we see that all of our choices of r have recovered at least $\mathcal{O}(G^{-1/2})$ convergence for the eigenfunction error. In Figure 6-6 we also see that all of our choices of r have recovered $\mathcal{O}(G^{-1})$ convergence for the eigenvalue error, however, choosing $\Delta = G^{-1/2}$ gives larger errors despite obtaining $\mathcal{O}(G^{-1})$ convergence. We also see that choosing $\Delta = G^{-1}$ and $\Delta = G^{-3/2}$ initially gives $\mathcal{O}(G^{-2})$ and $\mathcal{O}(G^{-3})$ convergence before "leveling off" to $\mathcal{O}(G^{-1})$ convergence once the errors have decayed to the levels of the method without smoothing. This final observation can also be

justified given the error dependence on Δ that we observed in Figures 6-3 and 6-4.

The results from Figures 6-5 and 6-6 both support our initial suggestion that we should choose $\Delta \lesssim G^{-1/2}$ to recover the convergence rates for the plane wave expansion method without smoothing. We also see that the errors with smoothing are consistently larger than or equal to the method without smoothing.

Now let us discuss the sampling method results. In Figures 6-7 and 6-8 we have plotted the errors for fixed $G = 2^8 - 1$ and varying sampling grid size, i.e. varying M. In both plots we see that the eigenvalue and eigenfunction errors decay with $\mathcal{O}(M^{-1})$, however, this decay rate is more pronounced for Model Problem 3. Note that we have only been able to plot results for particularly large M values because the method is unstable for smaller values of M. Also note that the eigenfunction errors in both of these figures stagnate for large M because the accuracy of the reference solutions is reached.

The fact that we observe errors that decay with $\mathcal{O}(M^{-1})$ suggests that we should choose $M \gtrsim N_f^{1/2}$ (recall that $N_f = 4G + 1$) to recover $\mathcal{O}(G^{-1/2})$ convergence for the eigenfunctions and $M \gtrsim N_f$ to recover $\mathcal{O}(G^{-1})$ convergence for the eigenvalues.

In Figures 6-9 and 6-10 we have experimented with choosing $M = N_f^r$ for different constants r. Although it is not very pronounced and we have been restricted by computational limitations, these figures are consistent with our conclusion that we should choose $M \gtrsim N_f$ to recover $\mathcal{O}(G^{-1})$ convergence in the eigenfunctions and eigenvalues. However, we also see that choosing larger M ($M = N_f^{3/2}$ or $M = N_f^2$) gives eigenfunction errors that are the same size as when exact Fourier coefficients are used. Unfortunately, we have not been able to plot enough points for the eigenvalue errors in Figure 6-10 to determine their convergence rates. Note that in Figures 6-9 and 6-10 our plots have again been limited in our choices of M since the method fails for M too small and is unfeasible for M large.

If we compare the Full 2D Problem (with sampling) with the Scalar 2D Problem (with sampling, see Section 4.4) then we see that the errors of both problems converge with $\mathcal{O}(M^{-1})$. It appears that convergence with M is independent of the regularity of the solution for these problems. Since convergence (with exact Fourier coefficients) is slower for the Full 2D Problem, we conclude that the sampling method is less harmful for the Full 2D Problem and it is easier to recover the optimal convergence rate.

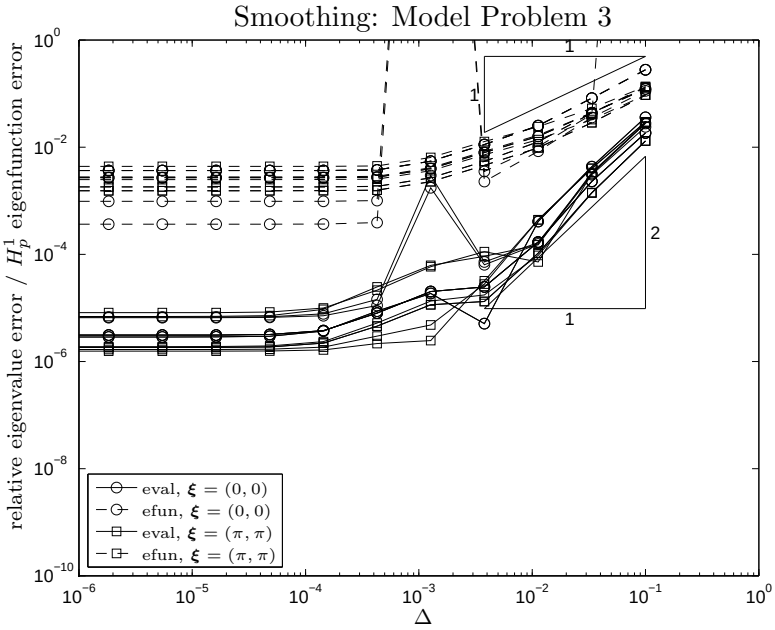

Figure 6-3: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. Δ for the 1st 5 eigenpairs of the plane wave expansion method with smoothing (G fixed) applied to Model Problem 3 for $\boldsymbol{\xi} = (0,0)$ and $\boldsymbol{\xi} = (\pi,\pi)$.

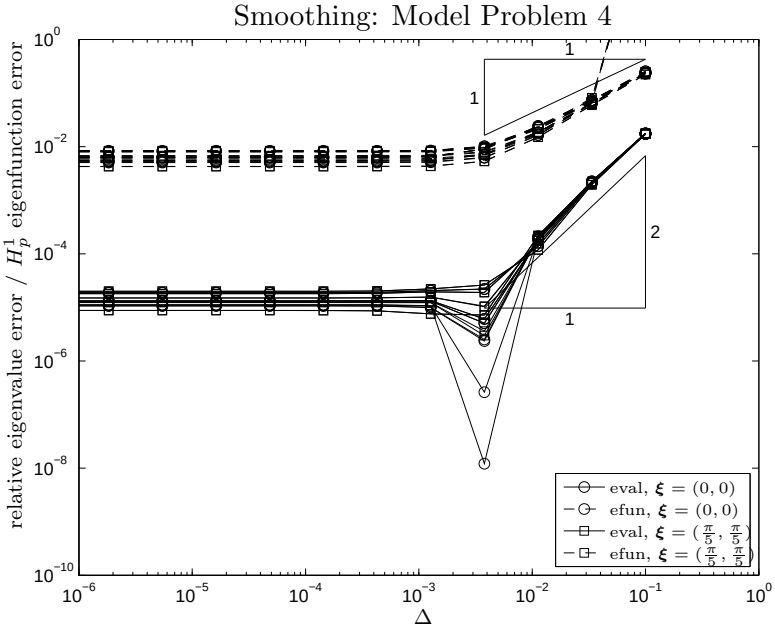

Figure 6-4: Plot of the relative eigenvalue error (eval) and the H_p^1 norm of the eigenfunction error (efun) vs. Δ for the 21st-30th eigenpairs of the plane wave expansion method with smoothing (G fixed) applied to Model Problem 4 for $\boldsymbol{\xi} = (0,0)$ and $\boldsymbol{\xi} = (\frac{\pi}{5},\frac{\pi}{5})$.

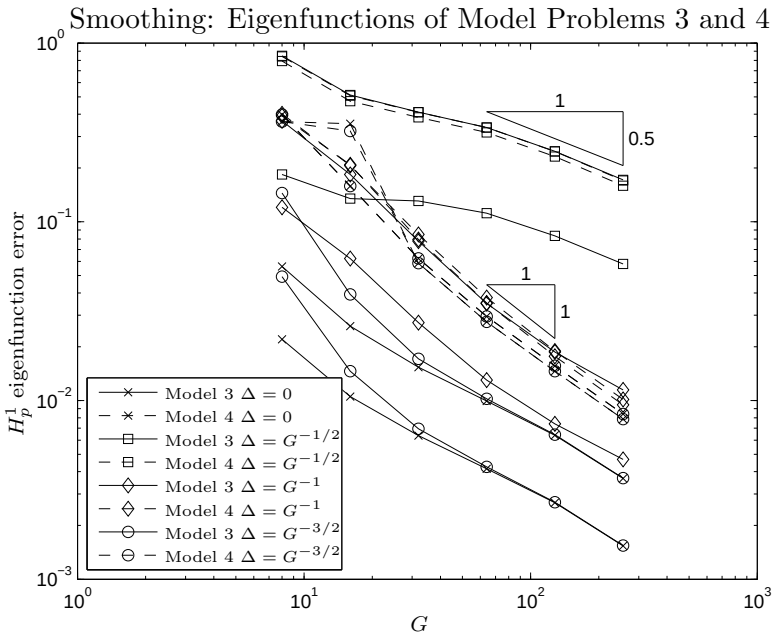

Figure 6-5: Plot of the H_p^1 norm of the error for the 1st eigenfunction vs. G for the plane wave expansion method with smoothing applied to Model Problems 3 and 4 for $\boldsymbol{\xi} = (0, 0)$, and $\boldsymbol{\xi} = (\pi, \pi)$ (for Model Problem 3) or $\xi = (\frac{\pi}{5}, \frac{\pi}{5})$ (for Model Problem 4).

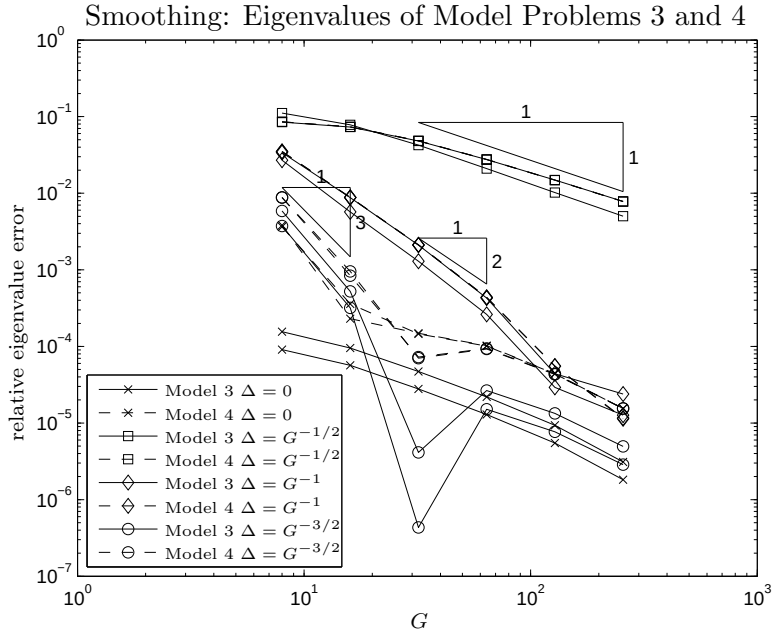

Figure 6-6: Plot of the relative error of the 1st eigenvalue vs. G for the plane wave expansion method with smoothing applied to Model Problems 3 and 4 for $\boldsymbol{\xi} = (0, 0)$, and $\boldsymbol{\xi} = (\pi, \pi)$ (for Model Problem 3) or $\boldsymbol{\xi} = (\frac{\pi}{5}, \frac{\pi}{5})$ (for Model Problem 4).

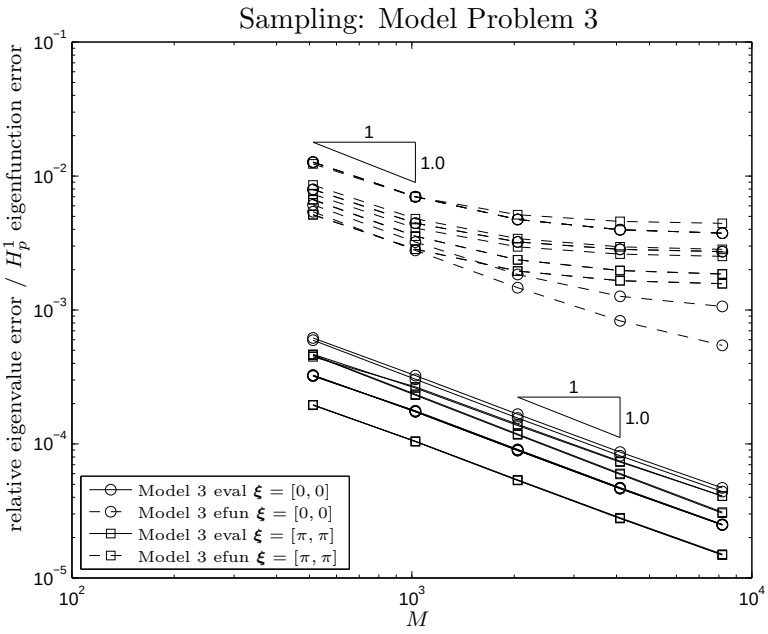

Figure 6-7: Plot of the relative eigenvalue error (eval) and the H_p^1 eigenfunction error (efun) vs. M for the 1st 5 eigenpairs of plane wave expansion method with sampling (fixed $G = 2^8 - 1$) applied to Model Problem 3 for $\xi = (0, 0)$, and $\xi = (\pi, \pi)$.

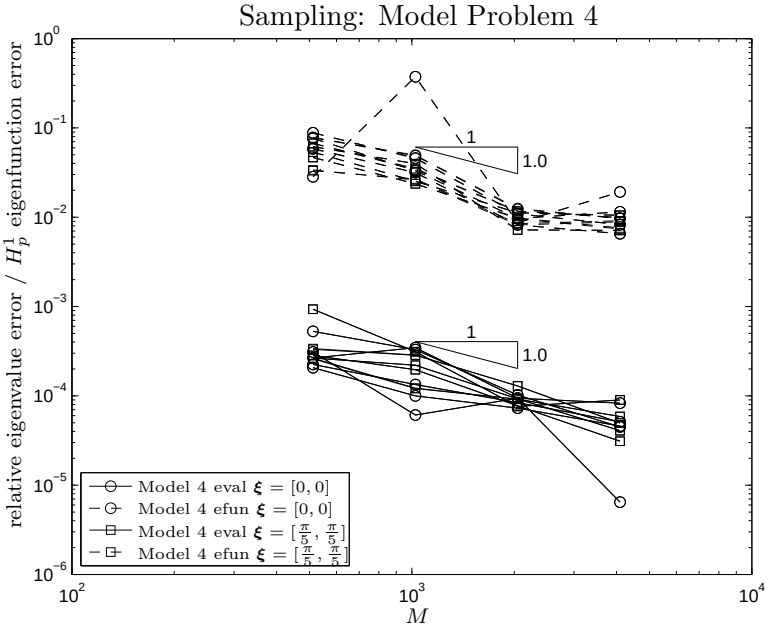

Figure 6-8: Plot of the relative eigenvalue error (eval) and the H_p^1 eigenfunction error (efun) vs. M for the 21st-30th eigenpairs of plane wave expansion method with sampling (fixed $G = 2^8 - 1$) applied to Model Problem 4 for $\xi = (0, 0)$, and $\xi = (\frac{\pi}{5}, \frac{\pi}{5})$.

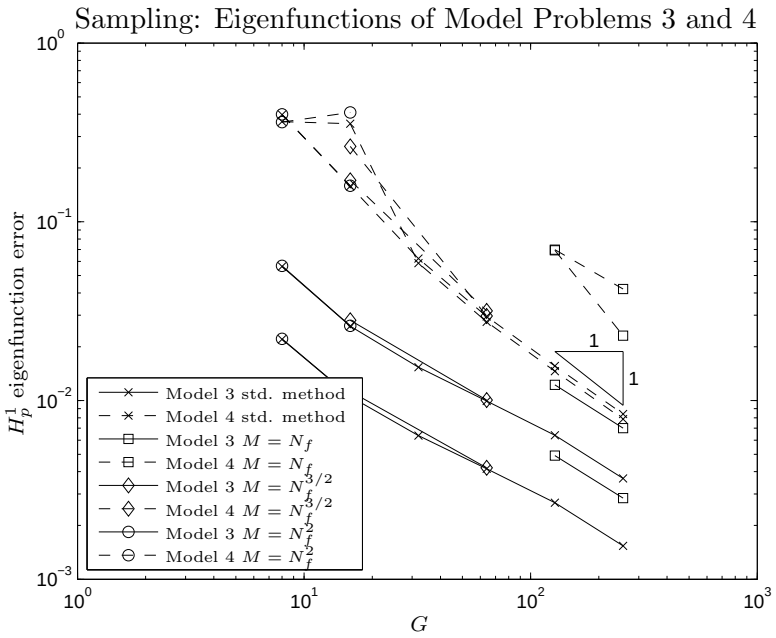

Figure 6-9: Plot of the H_p^1 norm of the error for the 1st eigenfunction vs. G for the plane wave expansion method with sampling applied to Model Problems 3 and 4 for $\boldsymbol{\xi} = (0, 0)$, and $\boldsymbol{\xi} = (\pi, \pi)$ (for Model Problem 3) or $\xi = (\frac{\pi}{5}, \frac{\pi}{5})$ (for Model Problem 4).

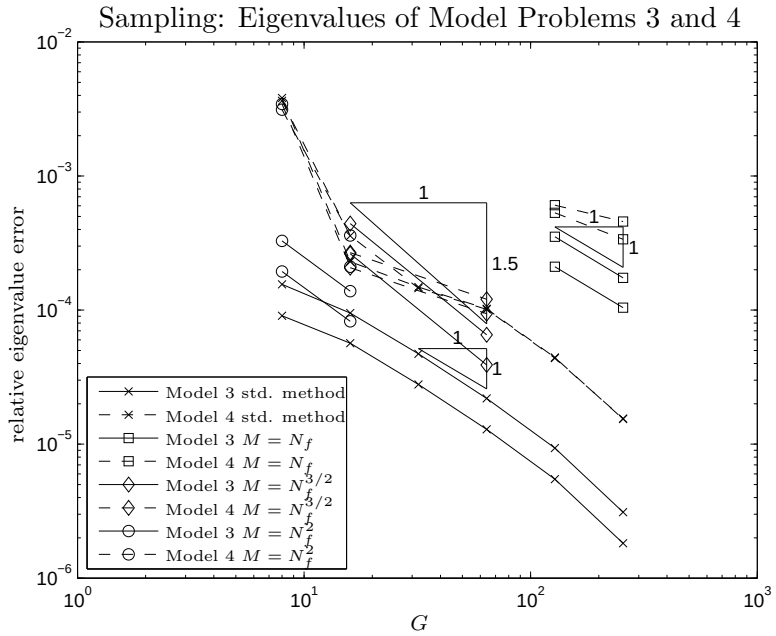

Figure 6-10: Plot of the relative error of the 1st eigenvalue vs. G for the plane wave expansion method with sampling applied to Model Problems 3 and 4 for $\boldsymbol{\xi} = (0, 0)$, and $\boldsymbol{\xi} = (\pi, \pi)$ (for Model Problem 3) or $\xi = (\frac{\pi}{5}, \frac{\pi}{5})$ (for Model Problem 4).

CHAPTER 7
CONCLUSIONS

In this chapter we briefly review the knowledge that we have gained on the plane wave expansion method, and its variations, and we put the success of the plane wave expansion method for the problems that we have studied into a wider perspective by making a comparison with the finite element method.

7.1 Review of the Plane Wave Expansion Method

In this thesis we have shown that the plane wave expansion method can be implemented efficiently for 4 different eigenvalue problems that come from photonic crystal fibres. We have observed and proved (or at least made significant progress towards proving) that the convergence of the plane wave expansion method depends directly on the regularity of each problem, which is limited because the coefficients of each problem are discontinuous. The limited regularity implies that the convergence of the method is *not* exponential (or superalgebraic). We have also shown that an attempt to recover superalgebraic convergence by smoothing the coefficients (the smoothing method) does not work because there is an additional error from smoothing. Also, since the plane wave expansion method requires the Fourier coefficients of the coefficients of each problem, we have presented an efficient method for approximating these Fourier coefficients (the sampling method) and we have shown how to recover the convergence rate of the plane wave expansion method with exact Fourier coefficients.

To apply the plane wave expansion method we first had to impose periodic boundary conditions (or periodic coefficients). For pure photonic crystals these arise naturally but for photonic crystal fibres they were imposed artificially by applying the supercell method. Although we have not proved any theoretical results for the error associated with the supercell method for any of our problems, we demonstrated for a particular ex-

ample in Figure 2-4 that the supercell method converges superalgebraically for isolated eigenvalues. Moreover, the essential spectrum can be accurately approximated with pure photonic crystal calculations without the supercell method. Further investigation into the supercell method could include computing more examples to confirm that the method converges superalgebraically (or even exponentially) for isolated eigenvalues in our other problems (not just the 1D TE Mode Problem) and trying to adapt the theory in [78] (where convergence of the supercell method is proven for 2D TE and TM Mode Problems) to our problems.

Applying the plane wave expansion method to each of our problems with periodic coefficients we obtained a matrix eigenproblem, which we solved using iterative techniques, for example, Implicitly Restarted Arnoldi and preconditioned CG or GMRES. Following [64] we used the Fast Fourier Transform (FFT) to obtain an efficient implementation for computing matrix-vector products with the system matrix A in $\mathcal{O}(N \log N)$ operations (N being the size of A) and we found that it is very easy to obtain an optimal preconditioner for A. These two implementation tricks are what make the plane wave expansion method competitive. For the 1D problems we solved matrix eigenproblems with $N \approx 5 \times 10^5$ in $\mathcal{O}(10^2)$ seconds and computed FFTs on vectors of length $2^{20} \approx 10^6$, whereas for the 2D problems we solved matrix eigenproblems where $N \approx 3 \times 10^6$ in $\mathcal{O}(10^3)$ seconds and computed 2D FFTs on matrices with dimension $2^1 2 \approx 4 \times 10^6$.

For the error analysis we considered the problems as spectral problems, applied the Floquet transform and obtained a variational eigenvalue problem. For all of our problems we developed regularity theory for the variational eigenvalue problems. For the 1D TE Mode Problem and the Scalar 2D Problem we discovered that the plane wave expansion method is a spectral Galerkin method and we were able to apply the theory from [6] to obtain error bounds in terms of the approximation error. We then used our regularity results to bound the approximation error for both the 1D TE Mode Problem and the Scalar 2D Problem, and we proved that the eigenfunction errors (measured in the H_p^1 norm) decay with $\mathcal{O}(G^{-3/2+\epsilon})$ for both of these problems (for all $\epsilon > 0$). We also proved that the eigenvalues decay at twice this rate. Using numerical examples we demonstrated (very clearly) that these error estimates are sharp (up to algebraic order).

For the 1D TM Mode Problem and the Full 2D Problem we could not show that the plane wave expansion method is a spectral Galerkin method and we could not apply the theory from [6] to complete an error analysis. Instead, we were limited to developing regularity results and bounding the approximation error. We showed that these problems had less regularity than the 1D TE Mode Problem and the Scalar 2D Problem and this was reflected in approximation error bounds that decayed more slowly, e.g. $\mathcal{O}(G^{-1/2+\epsilon})$ for all $\epsilon > 0$ for the eigenfunction errors (measured in the H_p^1

norm) in the case of the 1D TM Mode Problem. (For the Full 2D Problem we only managed to prove that the approximation error for the eigenfunctions is $\mathcal{O}(G^{-s})$ for *some* $s \geq 0$.) Although we did not manage to prove the stability of the plane wave expansion method for these two problems, we did observe stability in the numerical computations. Furthermore, we observed that the eigenvalues also converged at twice the rate of the eigenfunctions for these problems despite the matrix eigenproblem being non-symmetric.

It is suggested in [64] that replacing the discontinuous coefficients in each problem with smooth coefficients will recover superalgebraic (algebraic of arbitrary order) convergence for the plane wave expansion method. However, this introduces an additional error. We analysed the method that is used in [64] for the 1D TE Mode Problem and the Scalar 2D Problem and we proved that superalgebraic convergence to the "smooth problem" is obtained but that the additional error cancels any improvement. We devised an optimal strategy for balancing the smoothing error and the plane wave expansion error and this gave us a rate of convergence that was the same as the plane wave expansion method without smoothing. Numerical results confirmed our theory and showed that all but one of our estimates are sharp (up to algebraic order). The only exception is the dependence of the eigenvalue error on the amount of smoothing. We were only able to prove that this error decays at the same rate as the corresponding error in the eigenfunctions, but for some unknown reason we observe a slightly faster convergence rate (but not twice the rate of the eigenfunctions). We conclude that smoothing does not improve the plane wave expansion method for the 1D TE Mode Problem and the Scalar 2D Problem. We also computed numerical examples of smoothing for the 1D TM Mode Problem and the Full 2D Problem which agree with this conclusion.

The plane wave expansion method requires the Fourier coefficients of the coefficient functions to determine the entries of the matrix in the matrix eigenproblem. For 1D problems it is easy to construct an explicit formula for these Fourier coefficients, but in 2D it can easily be the case that the geometry of the photonic crystal fibre makes this task impossible. We examined the method that was used in [64] for approximating these Fourier coefficients. It is based on sampling the coefficient function on a uniform grid and then computing the FFT of the data to obtain approximate Fourier coefficients. We found (using theory for the 1D TE Mode Problem and the Scalar 2D Problem and numerical examples for all of the problems) that there is an additional error introduced by the sampling method, but the convergence rate with exact Fourier coefficients can be recovered if the sampling grid is chosen to have sufficiently small grid-spacing. For all of the problems we devised a strategy for choosing the optimal grid-spacing in relation to the size of the problem, and not surprisingly we found that it is easier to recover the (slower) convergence rate of the 1D TM Mode Problem and the Full 2D Problem than

the (faster) convergence rate of the 1D TE Mode Problem and the Scalar 2D Problem.

We also found that the plane wave expansion method with sampling is quite sensitive to the grid-spacing for the 1D TM Mode Problem and the Full 2D Problem. If it is chosen too large then the method fails. It is here that we see an opportunity for further investigation into some form of smoothing. If smoothing was applied *before* sampling then we might obtain a method that is not as sensitive to the grid-spacing. Therefore, we would recommend trying a different method for smoothing than the one we have considered in this thesis which acts more like a filter that is applied *after* sampling. For example, a different method for smoothing that might be more promising is considered in [40].

7.2 Comparison with the Finite Element Method

Now that we have reviewed our knowledge of the plane wave expansion method we would like to finish the thesis by comparing it with the finite element method. We will now explain why it compares favourably with the finite element method on a uniform grid.

When we apply both methods, the plane wave expansion method needs periodic boundary conditions, while the finite element method can be applied with any boundary conditions. This is not a disadvantage for the plane wave expansion method because the supercell method for imposing periodicity converges exponentially for the isolated eigenvalues and the essential spectrum can be calculated from the pure photonic crystal (that naturally has periodic coefficients).

For implementation, both methods give us a matrix eigenvalue problem to solve and we compare the two methods on two criteria, where there are differences: the cost of computing matrix-vector products; and the availability of an optimal preconditioner for solving linear systems. Matrix-vector products with the finite element method can be computed in $\mathcal{O}(N)$ operations (since the system matrix is sparse) whereas the plane wave expansion method requires $\mathcal{O}(N \log N)$ operations. This is a small advantage for the finite element method but the plane wave expansion method can use the simple preconditioner that we used in this thesis whereas the finite element method will require a more complicated multi-grid type preconditioner (unless K is large, in which case the finite element method can use the diagonal of the system matrix as a preconditioner).

For the convergence of these two methods, they are both restricted by the limited regularity of each of the problems that we have considered and therefore achieve similar convergence rates. However, the finite element method will need to use elements that have a higher order than piecewise linear elements in order to exploit the greater regularity of the 1D TE Mode Problem and the Scalar 2D Problem ($H_p^{5/2-\epsilon}$ for all $\epsilon > 0$). For the 1D TM Mode Problem and the Full 2D Problem the finite element

method will not need to use higher order elements because the regularity is not as high for these problems. Note that the methods may have different absolute errors despite converging at the same rate.

For 2D Problems, both methods can have difficulties representing complicated photonic crystal fibre structures. For the plane wave expansion method we require Fourier coefficients and we use the sampling method to approximate these, whereas there will be an additional error for the finite element method when the grid does not align with the interfaces of the discontinuous coefficients.

So far we have only considered the finite element method on a uniform grid and we see that neither method has a particular advantage over the other. Indeed, a case could be made that the plane wave expansion method is easier to implement and that "rough" calculations can more easily be made using it, but if we consider an adaptive finite element method, such as the method used in [31], with its plane wave equivalent, curvilinear coordinates, then we see that the finite element method gains an advantage.

Since the limited regularity of our problems is localised to the interface regions an adaptive finite element method will balance the limited regularity with a smaller grid size near the interfaces, resulting in a method that converges faster. Moreover, the grid will be more closely aligned with the interfaces to reduce error and multi-grid techniques can still be used to obtain an effective (if not optimal) preconditioner. The plane wave expansion method with curvilinear coordinates, on the other hand, does not have an optimal preconditioner since the derivative components from the operator are no longer confined to the diagonal of the matrix. An example of an adaptive finite element method applied to PCF problems is [31], where the 2D TE and TM Mode Problems are solved using *a posteriori* error estimation to refine the mesh.

To reiterate our final comparison conclusion, the plane wave expansion method compares favourably with the finite element method on a uniform grid but the adaptive finite element method has an advantage over the plane wave expansion method with curvilinear coordinates. However, an optimal preconditioner for the plane wave expansion method with curvilinear coordinates may be obtainable with further study.

APPENDIX A

EXTRA PROOFS

In this appendix we present some proofs that were not given in Chapter 3.

A.1 Lemma 3.3

The following is a proof of Lemma 3.3.

Proof. Suppose that Lemma 3.3 is not true. Then there exists a sequence $\phi_n \in \mathcal{D}(\mathbb{R}^d)$ such that

$$\frac{|\langle u, \phi_n \rangle|}{q_n(\phi_n)} =: c_n \to \infty \qquad \text{as } n \to \infty$$

where

$$q_n(\phi_n) = \sum_{|\alpha| \leq n} \max_{\mathbf{x} \in K} |D^\alpha \phi_n(\mathbf{x})|.$$

Now put

$$\psi_n = \frac{\phi_n}{c_n q_n(\phi_n)}.$$

Then $\psi_n \in \mathcal{D}(\mathbb{R}^d)$, $\operatorname{supp} \psi_n \subset K$ and

$$q_n(\psi_n) = \frac{1}{c_n} \to 0 \qquad \text{as } n \to \infty. \tag{A.1}$$

This implies that $\psi_n \to 0$ in $\mathcal{D}(\mathbb{R}^d)$ and so we have $\langle u, \psi_n \rangle \to 0$ as $n \to \infty$. But we also have (by the definition of ψ_n and c_n),

$$|\langle u, \psi_n \rangle| = \frac{1}{c_n q_n(\phi_n)} |\langle u, \phi_n \rangle| = 1 \qquad \forall n \in \mathbb{N}.$$

This is a contradiction. □

A.2 Piecewise Continuous Functions

The following proof is a proof of Lemma 3.38.

Proof. We present the proof for $d \geq 2$. The $d = 1$ proof is similar and easier. The proof is given in two steps:

1. Show $|\widehat{f}(\mathbf{k})| \leq C_{m,u}(1 + |\mathbf{k}'|)^{-m}(1 + |k_d|)^{-1}$ for all $\mathbf{k} \in \mathbb{R}^d$ and for every $m \in \mathbb{N}$.

2. Show $\|f\|^2_{H^s(\mathbb{R}^d)} = \int_{\mathbb{R}^d}(1 + |\mathbf{k}|^2)^s|\widehat{f}(\mathbf{k})|^2 d\mathbf{k} < \infty$ for $s < 1/2$.

Step 1. Let $\mathbf{k} \in \mathbb{R}^d$ and recall the notation: $\mathbf{k}' = (k_1, k_2, \ldots, k_{d-1})$. Let k_j denote the element of $\mathbf{k}'$ with maximum absolute value and define $U := |\operatorname{supp} u|$ and $U' := |\operatorname{supp} u(\mathbf{x}', 0)|$. We will need the following inequality,

$$|k_j|^m \leq |\mathbf{k}'|^m \leq (d-1)^{m/2}|k_j|^m \qquad \forall m > 0. \tag{A.2}$$

We begin with the definition of $\widehat{f}(\mathbf{k})$ and integrate by parts to get the following equalities with $p \in \mathbb{N} \cup \{0\}$.

$$\begin{aligned}
\widehat{f}(\mathbf{k}) &= \int_{\mathbb{R}^d} \mathrm{e}^{-i2\pi\mathbf{k}\cdot\mathbf{x}} f(\mathbf{x}) d\mathbf{x} \\
&= \int_{x_d<0} \mathrm{e}^{-i2\pi\mathbf{k}\cdot\mathbf{x}} u(\mathbf{x}) d\mathbf{x} \\
&= \frac{1}{(i2\pi k_j)^p} \int_{x_d<0} \mathrm{e}^{-i2\pi\mathbf{k}\cdot\mathbf{x}} D_j^p u(\mathbf{x}) d\mathbf{x} \\
&= \frac{1}{(i2\pi k_j)^p} \frac{1}{i2\pi k_d} \left(\int_{x_d<0} \mathrm{e}^{-i2\pi\mathbf{k}\cdot\mathbf{x}} D_d \partial_j^p u(\mathbf{x}) d\mathbf{x} - \int_{x_d=0} \mathrm{e}^{-i2\pi\mathbf{k}'\cdot\mathbf{x}'} (D_j^p u)|_{x_d=0} d\mathbf{x}' \right)
\end{aligned}$$

Using these equalities and (A.2) we get the following

$$|\widehat{f}(\mathbf{k})| \leq \begin{cases} U\|u\|_{L^\infty(\mathbb{R}^d)} \\ \frac{U}{(2\pi|k_j|)^p}\|D_j^p u\|_{L^\infty(\mathbb{R}^d)} \\ \frac{1}{(2\pi)^{p+1}|k_j|^p|k_d|}\left(U\|D_d D_j^p u\|_{L^\infty(\mathbb{R}^d)} + U'\|D_j^p u|_{x_d=0}\|_{L^\infty(\mathbb{R}^{d-1})}\right) \end{cases} \tag{A.3}$$

for $p \in \mathbb{N} \cup \{0\}$. Now let $m \in \mathbb{N}$ and consider $(1 + |\mathbf{k}'|)^m(1 + |k_d|)|\widehat{f}(\mathbf{k})|$ for different cases of $|\mathbf{k}'|$ and $|k_d|$.
Case 1: If $|\mathbf{k}'| < 1$ and $|k_d| < 1$, then

$$\begin{aligned}
(1 + |\mathbf{k}'|)^m(1 + |k_d|)|\widehat{f}(\mathbf{k})| &\leq 2^{m+1}|\widehat{f}(\mathbf{k})| \\
&\leq 2^{m+1}U\|u\|_{L^\infty(\mathbb{R}^d)} \qquad \text{by (A.3).}
\end{aligned}$$

Case 2: If $|\mathbf{k}'| > 1$ and $|k_d| < 1$, then

$$\begin{aligned}(1+|\mathbf{k}'|)^m(1+|k_d|)|\widehat{f}(\mathbf{k})| &\le 2^{m+1}|\mathbf{k}'|^m|\widehat{f}(\mathbf{k})| \\ &\le 2^{m+1}(d-1)^{m/2}|k_j|^m|\widehat{f}(\mathbf{k})| \qquad \text{by (A.2)} \\ &\le \frac{2(d-1)^{m/2}U}{\pi^m}\|D_j^m u\|_{L^\infty(\mathbb{R}^d)} \qquad \text{by (A.3) with } p=m.\end{aligned}$$

Case 3: If $|\mathbf{k}'| < 1$ and $|k_d| > 1$, then

$$\begin{aligned}(1+|\mathbf{k}'|)^m(1+|k_d|)|\widehat{f}(\mathbf{k})| &\le 2^{m+1}|k_d||\widehat{f}(\mathbf{k})| \\ &\le \frac{2^m}{\pi}\left(U\|D_d u\|_{L^\infty(\mathbb{R}^d)} + U'\|u|_{x_d=1}\|_{L^\infty(\mathbb{R}^{d-1})}\right) \quad \text{by (A.3) with } p=0.\end{aligned}$$

Case 4: If $|\mathbf{k}'| > 1$ and $|k_d| > 1$, then

$$\begin{aligned}(1+|\mathbf{k}'|)^m(1+|k_d|)|\widehat{f}(\mathbf{k})| &\le 2^{m+1}|\mathbf{k}'|^m|k_d||\widehat{f}(\mathbf{k})| \\ &\le 2^{m+1}(d-1)^{m/2}|k_j|^m|k_d||\widehat{f}(\mathbf{k})| \qquad \text{by (A.2)} \\ &\le \frac{(d-1)^{m/2}}{\pi^{m+1}}\left(U\|D_d D_j^m u\|_{L^\infty} + U'\|D_j^m u|_{x_d=0}\|_{L^\infty}\right) \quad \text{by (A.3) with } p=m.\end{aligned}$$

Since $u \in C_0^\infty(\mathbb{R}^d)$, the right-hand-sides of Cases 1-4 are all bounded by constants that depend on m, u and d and we have completed Step 1.
Step 2. For any $\mathbf{k} \in \mathbb{R}^d$ we get

$$1+|\mathbf{k}|^2 = 1+|\mathbf{k}'|^2+|k_d|^2 \le (1+|\mathbf{k}'|)^2(1+|k_d|)^2 \tag{A.4}$$

$$\begin{aligned}\|f\|_{H^s(\mathbb{R}^d)}^2 &= \int_{\mathbb{R}^d}(1+|\mathbf{k}|^2)^s|\widehat{f}(\mathbf{k})|^2 d\mathbf{k} \\ &\le C_{m,u}^2\int_{\mathbb{R}^d}\frac{(1+|\mathbf{k}|^2)^s}{(1+|\mathbf{k}'|)^{2m}(1+|k_d|)^2}d\mathbf{k} \qquad \forall m\in\mathbb{N} \text{ by Step 1} \\ &= C_{m,u}^2\int_{\mathbb{R}^d}\frac{(1+|\mathbf{k}'|)^{2s}(1+|k_d|)^{2s}}{(1+|\mathbf{k}'|)^{2m}(1+|k_d|)^2}d\mathbf{k} \qquad \forall m\in\mathbb{N} \text{ by (A.4)} \\ &= C_{m,u}^2\underbrace{\left(\int_{\mathbb{R}^{d-1}}(1+|\mathbf{k}'|)^{2s-2m}d\mathbf{k}'\right)}_{I_1}\underbrace{\left(\int_{\mathbb{R}}(1+|k_d|)^{2s-2}dk_d\right)}_{I_2} \quad \forall m\in\mathbb{N}\end{aligned}$$

The term I_1 is bounded by choosing m sufficiently large and the term I_2 is bounded provided $2s-2<-1$, or equivalently, if $s<1/2$. This completes the proof. □

A.3 Triangle Inequality for Gap Between Subspaces

The *gap* between two subspaces of a Hilbert space (Definition 3.64) obeys the triangle inequality, Lemma 3.65. Here is the proof for Lemma 3.65.

Proof. Let X, Y and Z be three closed subspaces of a Hilbert space. The proof has three steps.

1. Since $\{y \in Y : \|y\| = 1\} \subset \{y \in Y : \|y\| \leq 1\}$,

$$\sup_{y\in Y, \|y\|\leq 1} \operatorname{dist}(y, Z) \geq \sup_{y\in Y, \|y\|=1} \operatorname{dist}(y, Z). \tag{A.5}$$

Conversely, for each $0 \neq y \in Y$, with $\|y\| \leq 1$, define $\hat{y} = \frac{y}{\|y\|}$. Then

$$\operatorname{dist}(\hat{y}, Z) = \inf_{z\in Z} \|\hat{y} - z\| = \frac{1}{\|y\|} \inf_{z'\in Z} \|y - z'\| = \frac{1}{\|y\|} \operatorname{dist}(y, Z) \geq \operatorname{dist}(y, Z)$$

since $\|y\| \leq 1$. Therefore

$$\sup_{y\in Y, \|y\|\leq 1} \operatorname{dist}(y, Z) \leq \sup_{y\in Y, \|y\|=1} \operatorname{dist}(y, Z). \tag{A.6}$$

Combining (A.5) and (A.6) we get

$$\sup_{y\in Y, \|y\|\leq 1} \operatorname{dist}(y, Z) = \sup_{y\in Y, \|y\|=1} \operatorname{dist}(y, Z). \tag{A.7}$$

2. For $x \in X$, $\|x\| = 1$, since $\{y \in Y : \|y\| \leq 1\} \subset Y$,

$$\inf_{y\in Y, \|y\|\leq 1} \|x - y\| \geq \inf_{y\in Y} \|x - y\| \tag{A.8}$$

Conversely, let y_x be the projection of x onto Y with respect to the inner product on our Hilbert space, $(x - y_x, y) = 0$ for all $y \in Y$. Then, using the definition of y_x, Cauchy-Schwarz and that $\|x\| = 1$, we get

$$\|y_x\|^2 = (y_x, y_x) = (x, y_x) \leq \|x\|\|y_x\| = \|y_x\|.$$

Therefore $\|y_x\| \leq 1$. Also, Pythagorus gives us

$$\|x - y\|^2 = \|x - y_x\|^2 + \|y_x - y\|^2 \qquad \forall y \in Y$$

which implies

$$\|x - y\| \geq \|x - y_x\| \qquad \forall y \in Y.$$

Therefore,

$$\inf_{y\in Y}\|x-y\| \geq \|x-y_x\| \geq \inf_{y'\in Y,\|y'\|\leq 1}\|x-y'\| \tag{A.9}$$

Combining (A.8) and (A.9) we get, for $x \in X$ with $\|x\| = 1$,

$$\inf_{y\in Y,\|y\|\leq 1}\|x-y\| = \inf_{y\in Y}\|x-y\| \tag{A.10}$$

3. Let $x \in X$ with $\|x\| = 1$. Then

$$\begin{aligned}
\operatorname{dist}(x,Z) &= \inf_{z\in Z}\|x-z\| \\
&\leq \|x-y\| + \inf_{z\in Z}\|y-z\| && \forall\, y\in Y, \|y\|\leq 1 \\
&= \|x-y\| + \operatorname{dist}(y,Z) \\
&\leq \|x-y\| + \sup_{y'\in Y,\|y'\|\leq 1}\operatorname{dist}(y',Z) \\
&= \|x-y\| + \sup_{y'\in Y,\|y'\|=1}\operatorname{dist}(y',Z) && \text{by (A.7)} \\
&= \|x-y\| + \delta(Y,Z) && \forall\, y\in Y, \|y\|\leq 1.
\end{aligned}$$

Taking the infimum over $y \in Y$ with $\|y\| \leq 1$ we get

$$\begin{aligned}
\operatorname{dist}(x,Z) &\leq \inf_{y\in Y,\|y\|\leq 1}\|x-y\| + \delta(Y,Z) \\
&= \inf_{y\in Y}\|x-y\| + \delta(Y,Z) && \text{by (A.10)} \\
&= \operatorname{dist}(x,Y) + \delta(Y,Z).
\end{aligned}$$

The result follows by taking the supremum over $x \in X$ with $\|x\| = 1$.

□

BIBLIOGRAPHY

[1] Abramowitz, M. & Stegun, I.A., *Handbook of Mathematical Functions*, 1965.

[2] Adams, R.A. & Fournier, J.J.F., *Sobolev Spaces 2nd Edition*, 2003.

[3] Ashcroft, N.W. & Mermin, N.D. *Solid State Physics*, 1976.

[4] Asplund, E. & Bungart, L., *A First Course in Integration*, 1966.

[5] Axmann, W. & Kuchment, P., An efficient finite element method for computing spectra of photonic and acoustic band-gap materials, *Journal of Computational Physics*, 150, pp. 468-481, 1999.

[6] Babuska, I. & Osborn J., *Eigenvalue Problems*, 1991.

[7] Birks, T.A., Bird, D.M. et. al., Scaling laws and vector effects in bandgap-guiding fibres, *Optics Express*, 12, pp. 69-74, 2004.

[8] Cao, Y., Hou, Z. & Liu, Y., Convergence problem of plane-wave expansion method for photonic crystals, *Physics Letters A*, 327, pp. 247-253, 2004.

[9] Ciarlet, P.G., *The Finite Element Method for Elliptic Problems*, 1978.

[10] Cooley, J. & Tukey, J., An algorithm for the machine calculation of complex Fourier series, Mathematics of Computation, 19, pp. 297-301, 1965.

[11] Dangui, V., Digonnet, M.J.F. & Kino, G.S., A fast and accurate numerical tool to model the modal properties of photonic-bandgap fibers, *Optics Express*, 14, pp. 2979-2993, 2006.

[12] Dautray, R. & Lions, J.L., *Mathematical Analysis and Numerical Methods for Science and Technology, Volume I*, 1990.

[13] Demmel, J.W., *Applied Numerical Linear Algebra*, 1997.

[14] Dhia, A.S.B.B. & Gmati, N., Spectral approximation of a boundary condition for an eigenvalue problem, *SIAM Journal on Numerical Analysis*, 32, pp. 1263-1279, 1995.

[15] Dobson, D.C., An efficient method for band structure calculations in 2D photonic crystals, *Journal of Computational Physics*, 149, pp. 363-376, 1999.

[16] Duoandikoetxea, J., *Fourier Analysis*, 2001.

[17] Eastham, M.S.P., *The Spectral Theory of Periodic Differential Operators*, 1973.

[18] Elschner, J., *Singular Ordinary Differential Operators and Pseudodifferential Equations*, 1985.

[19] Elschner, J., Hinder, R. et al., Existence, uniqueness and regularity for solutions of the conical diffraction problem, *Mathematical Models and Methods in Applied Sciences*, 10, pp. 317-341, 2000.

[20] Elschner, J. & Schmidt, G., Conical diffraction by periodic structures: variation of interfaces and gradient formulas, *Mathematische Nachrichten*, 252, pp. 24-42, 2003.

[21] Evans, L.C., *Partial Differential Equations*, 1998.

[22] Feit, M.D. & Fleck Jr., J.A., Computation of mode properties in optical fiber waveguides by a propagating beam method, *Applied Optics*, 19, pp. 1154-1164, 1980.

[23] Figotin, A. & Godin, Y., The computation of spectra of some 2D photonic crystals, *Journal of Computational Physics*, 136, pp. 585-598, 1997.

[24] Figotin, A. & Klein, A., Localization of classical waves II: electromagnetic waves, *Communications in Mathematical Physics*, 184, pp. 411-441, 1997.

[25] Figotin, A. & Kuchment, P., Band-gap structure of spectra of periodic dielectric and acoustic media. I. scalar model, *SIAM Journal on Applied Mathematics*, 56, pp. 68-88, 1996.

[26] Figotin, A. & Kuchment, P., Band-gap structure of spectra of periodic dielectric and acoustic media. II. two-dimensional photonic crystals, *SIAM Journal on Applied Mathematics*, 56, pp. 1561-1620, 1996.

[27] Figotin, A. & Kuchment,P., Spectral properties of classical waves in high-contrast periodic media, *SIAM Journal on Applied Mathematics*, 58, pp. 683-702, 1998.

[28] Filonov, N., Gaps in the spectrum of the Maxwell operator with periodic coefficients, *Communications in Mathematical Physics*, 240, pp. 161-170, 2003.

[29] Fliss, S., Joly, P. & Li J.R., Exact boundary conditions for locally perturbed 2D-periodic plane, *Proceedings of Waves 2007*, pp. 495-497, 2007.

[30] Frigo, M. & Johnson, S.G., The design and implemenation of FFTW3, *Proceedings of the IEEE*, 93(2), pp. 216-231, 2005.

[31] Giani, S., *Convergence of Adaptive Finite Element Methods for Elliptic Eigenvalue Problems with Application to Photonic Crystals*, PhD Thesis, University of Bath, 2008.

[32] Grisvard, P., *Singularities in Boundary Value Problems*, 1992.

[33] Guan, N., Habu, S., et al., Boundary element method for analysis of holey optical fibers, *Journal of Lightwave Technology*, 21, pp. 1787-1792, 2003.

[34] Guo, S. & Albin, S., Simple plane wave implementation for photonic crystal calculations, *Optics Express*, 11, pp. 167-175, 2003.

[35] Hackbusch, W., *Elliptic Differential Equations*, 1992.

[36] Hardy, G.H. & Rogosinski, W.W., *Fourier Series*, 1956.

[37] Hislop, P.D. & Sigal, I.M., Introduction to Spectral Theory with Applications to Schrödinger Operators, 1996.

[38] Ho, K.M., Chan, C.T. & Soukoulis, C.M., Existence of a photonic gap in periodic dielectric structures, *Physical Review Letters*, 65, pp. 3152-3155, 1990.

[39] Joannopoulos, J.D., Meade, R.D. & Winn, J.N., *Photonic Crystals: Molding the Flow of Light*, 1995.

[40] Johnson, S.G. & Joannopoulos, J.D., Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis, *Optics Express*, 8, pp. 173-190, 2000.

[41] John, S., Strong localization of photons in certain disordered dielectric superlattices, *Physical review letters*, 58, pp.2486-2489, 1987.

[42] Kato, T., *Perturbation theory for linear operators*, 1966.

[43] Kelley, C.T., *Iterative Methods for Linear and Nonlinear Equations*, 1995.

[44] Kuchment, P., *Floquet Theory for Partial Differential Equations*, 1993.

[45] Kuchment, P., The mathematics of photonic crystals. Ch 7. in *Mathematical Modelling in Optical Science, Frontiers in Applied Mathematics*, 22, pp. 207-272, 2001.

[46] Kuchment, P., On some spectral problems of mathematical physics, in *Contempory Mathematics, Partial Differential Equations and Inverse Problems*, 362, pp. 241-276, 2003.

[47] Kuchment, P. & Ong, B.S., On guided waves in photonic crystal waveguides, in *Comtempory Mathematics, Waves in Periodic and Random Media*, 339, pp. 105-115, 2004.

[48] Knight, J.C., Photonic crystal fibres, *Nature*, 424, pp. 847-851, 2003.

[49] Kunz, K.S. & Luebbers, R.J., *The Finite Difference Time Domain Method for Electromagnetics*, 1993.

[50] Lax, P.D., *Functional Analysis*, 2002.

[51] Lehoucq, R.B., Sorensen, D.C. & Yang, C., *ARPACK Users' Guide*, 1998.

[52] Lions, J.L. & Magenes, E., *Non-Homogeneous Boundary Value Problems and Applications*, 1972.

[53] Meade, R.D., Rappe, A.M. et al., Accurate theoretical analysis of photonic band-gap materials, *Physical Review B*, 48, pp. 8434-8437, 1993.

[54] McLean, W., *Strongly Elliptic Systems and Boundary Integral Equations*, 2000.

[55] Merzbacher, E., *Quantum Mechanics*, 1961.

[56] Mogilevtsev, D., Birks, T.A. & Russell, P.St.J., Localized function method for modeling defect modes in 2-D photonic crystals, *Journal of Lightwave Technology*, 17, pp. 2078-2081, 1999.

[57] Monk, P., *Finite Element Methods for Maxwell's Equations*, 2003.

[58] Monro, T.M., Richardson, D.J. et al., Modeling large air fraction holey optical fibers, *Journal of Lightwave Technology*, 18, pp. 50-56, 2000.

[59] Morame, A., The absolute continuity of the spectrum of Maxwell operator in periodic media, *Journal of Mathematical Physics*, 41, pp. 7099-7108, 2000.

[60] Móricz, F., Pointwise behaviour of double Fourier series of functions of bounded variation, *Monatshefte für Mathematik*, 148, pp. 51-59, 2006.

[61] Payne, M.C., Teter, M.P. et al., Iterative minimization techniques for *ab initio* total energy calculations: molecular dynamics and conjugate gradients, *Reviews of Modern Physics*, 64(4), pp. 1045-1097, 1992.

[62] Pearce, G.J., Pottage, J.M. et al., Hollow-core PCF for guidance in the mid to far infra-red, *Optics Express*, 13, pp. 6937-6945, 2005.

[63] Pearce, G.J., Hedley, T.D. & Bird, D.M., Adaptive curvilinear coordinates in a plane-wave solution of Maxwell's equations in photonic crystals, *Physical Review B*, 71, pp. 195108(10), 2005.

[64] Pearce, G.J., *Plane-wave methods for modelling photonic crystal fibre*, PhD Thesis, University of Bath, 2006.

[65] Petzoldt, M., *Regularity and error estimators for elliptic problems with discontinuous coefficients*, PhD Thesis, FU Berlin, http://www.diss.fu-berlin.de/diss, 2001.

[66] Pottage, J.M., Bird, D.M. et al., Robust photonic band gaps for hollow core guidance in PCF made from high index glass, *Optics Express*, 11, pp. 2854-2861, 2003.

[67] Price, J.F. & Sloan, I.H., Pointwise convergence of multiple Fourier series: sufficient conditions and an application to numerical integration, *Journal of Mathematical Analysis and Applications*, 169, pp. 140-156, 1992.

[68] Qiu, M., Analysis of guided modes in photonic crystal fibres using the finite-difference time-domain method, *Microwave and Optical Technology Letters*, 30, pp. 327-330, 2001.

[69] Reed, M. & Simon, B., *Methods of Modern Mathematical Physics IV Analysis of Operators*, 1978.

[70] Russell, P.St.J., Photonic Crystal Fibers, *Science*, 299, pp. 358-362, 2003.

[71] Saitoh, K. & Koshiba, M., Full-vectorial imaginary-distance beam propagation method based on a finite element scheme: Application to photonic crystal fibres, *IEEE Journal of Quantum Electronics*, 38, pp. 927-933, 2002.

[72] Saranen, J. & Vainikko, G., *Periodic Integral and Pseudodifferential Equations with Numerical Approximation*, 2002.

[73] Saad, Y., *Numerical Methods for Large Eigenvalue Problems*, 1992.

[74] Saad, Y., *Iterative methods for sparse linear systems*, 2003.

[75] Shen, L. & He, S., Analysis for the convergence problem of the plane-wave expansion method for photonic crystals, *Journal for the Optical Society of America A*, 19, pp. 1021-1024, 2002.

[76] Snyder, A.W. & Love, J.D. *Optical Waveguide Theory*, 1983.

[77] Sorensen, D.C., Implicity application of polynomial filters in a k-step Arnoldi method, *SIAM Journal on Matrix Analysis and Applications*, 13, pp. 357-385, 1992.

[78] Soussi, S., Convergence of the supercell method for defect modes calculations in photonic crystals, *SIAM Journal on Numerical Analysis*, 43, pp. 1175-1201, 2005.

[79] Sözüer, H.S. & Haus, J.W., Photonic bands: convergence problems with the plane-wave method, *Physical Review B*, 45, pp. 13962-13973, 1992.

[80] Stein, E.M. & Weiss, G., *Introduction to Fourier Analysis in Euclidean Spaces*, 1971.

[81] Strang, G., *Introduction to Applied Mathematics*, 1986.

[82] Taflove, A. & Hagness, S.C., *Computational electrodynamics: The finite-difference time-domain method (3rd ed.)*, 2005.

[83] Trefethen, L.N. & Bau, D., *Numerical Linear Algebra*, 1997.

[84] Van Loan, C., *Computational Frameworks for the Fast Fourier Transform*, 1992.

[85] Vanselow, R., Convergence analysis for the full-upwind finite volume solution of a convection-diffusion problem, *Journal of Mathematical Analysis and Applications*, 264, pp. 423-449, 2001.

[86] Wang, X., Lou, J. et al., Modeling of PCF with multiple reciprocity boundary element method, *Optics Express*, 12, pp. 961-966, 2004.

[87] Watkins, D.S., *Fundamentals of Matrix Computations (2nd Edition.)*, 2002.

[88] White, T.P., McPhedran, R.C. et al., Confinement losses in microstructured optical fibers, *Optics Letters*, 26, pp. 1660-1662, 2001.

[89] White, T.P., Kuhlmey, B.T. et al., Multipole method for microstructured optical fibres. I. Formulation, *Journal for the Optical Society of America*, 19, pp. 2322-2330, 2002.

[90] Yablonovitch, E., Inhibited Spontaneous Emission in Solid-State Physics and Electronics, *Physical review letters*, 58, pp. 2059-2062, 1987.

[91] Yeh, P. & Yariv, A., Theory of Bragg fiber, *Journal of the Optical Society of America*, 68, pp. 1196-1201, 1978.